Progress testing : the utility of an assessment concept

Progress Testing

The Utility of an Assessment Concept

Omslagontwerp:
Jan-Henk de Vree

ISBN 90-367-1805-8

Druk: Stichting Drukkerij C. Regenboog, Groningen, The Netherlands

Progress Testing

The Utility of an Assessment Concept

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Maastricht,
op gezag van de Rector Magnificus
Prof. dr. A.C. Nieuwenhuijzen Kruseman,
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen op
vrijdag 25 april 2003 om 14.00 uur

door

Bas Henk Verhoeven
geboren op 10 januari 1970 te Oisterwijk

Promotores:

Prof. dr. C.P.M. van der Vleuten
Prof. dr. A.J.J.A. Scherpbier
Prof. dr. W.H.F.W. Wijnen

Beoordelingscommissie:

Prof. dr. M.F. von Meyenfeldt (voorzitter)
Dr. Y. van Leeuwen
Dr. ir. A.M.M. Muijtjens
Prof. dr. R. van Schilfgaarde (Rijksuniversiteit Groningen)
Prof. dr. R.P. Zwierstra (Rijksuniversiteit Groningen)

Paranimfen:

Drs. G.M. Verwijnen
Drs. W.H. Roerdink

Progress Testing

The Utility of an Assessment Concept

Toen ik langs het tuinpad van m'n vader
De hoge bomen nog zag staan
Ik was een kind, hoe kon ik weten
Dat dat voorgoed voorbij zou gaan
(Het Dorp, Wim Sonneveld)

Voor mijn vader

Papa, why do you play
All the same old songs
Why do you sing
With the melody

Cause down on the street
Somethings goin' on
There's a new beat
And a brand new song

He said
In my life, there was so much anger
Still I have no regrets
Just like you, I was such a rebel
So dance your own dance, and never forget

N'oubliez jamais
I heard my father say
Every generation has it's way
A need to disobey

N'oubliez jamais
It's in your destiny
A need to disagree
When rules get in the way
N'oubliez jamais

(N'oubliez Jamais, Joe Cocker)

Contents

Chapter 1

Introduction

Assessment

Tests of educational achievement are an ever fascinating and rewarding topic for discussion wherever they are applied. Keeping one's ear to the ground during secondary school breaks, in a students' bar or in the canteen of the driving school one can enjoy many provocative and heated debates about content, relevance, method, fairness and standards of assessment. Also for many teachers, curriculum designers and educationalists, examinations are a perpetual source of controversy. It is an area where tradition, personal values and experiences tend to dominate the debate.[1,2] On the other hand the number of scientific publications on assessment and assessment methods is huge and assessment still remains one of the major topics in the literature on education. Many different assessment methods have been developed and described during the last decades. One of them is the progress test which is a longitudinal final objective assessment method. This dissertation focuses on the utility of progress testing as an assessment concept in undergraduate medical education.

Structure of the dissertation
In the first chapter some background information and a brief history of the progress test will be presented. In the second chapter a model for evaluating the utility of assessment methods will be described and the research questions that are addressed in this dissertation stated. Chapters three through eleven report on the studies conducted to answer the research questions. Chapter twelve will synthesize the findings and expound the answers to the research questions. Finally, the dissertation will be summarized in English and Dutch.

Note
This dissertation is based on nine articles most of which have been published. Each article was written to be read independently. This means that some repetition of information across chapters was inevitable.

Background

Knowledge and competence

Epstein and Hundert define medical professional competence as "the habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values, and reflection in daily practice for the benefit of the individual and community being served" and claim it builds on a foundation of basic clinical skills, scientific knowledge, and moral development.[3] These competences appear to share many aspects and are not independent of each other. Assessments of different aspects of competence have been found to be strongly correlated.[4] Competence appears to be strongly associated with knowledge and experience in a specific content area. Achieving competence in one area is not a good predictor of competence in another, even if the areas are closely related.[5] Ability is not easily transferred from one situation or problem to another and therefore it is possible for one person to function at several cognitive levels, depending on the problem at hand.[2] Many studies on clinical reasoning have confirmed that domain-specific knowledge is a central factor in medical competence. Although expert reasoning is strongly connected with knowledge, it is not simply the knowledge itself that determines expertise, but the way it is stored, organized, retrieved and used. Medical expertise appears to be based upon a doctors' well-developed, highly structured and reshapeable knowledge networks.[6-14] Developing a comprehensive and functional knowledge base is therefore essential for medical students enabling them to become medical experts. In this respect assessment of knowledge is essential.

Lifelong and active learning

Knowledge dates fast and the increase of knowledge within even a rather narrow field is daunting. To be able to keep up to date in order to provide patients with optimal care and master the information overload, students and physicians have to be lifelong learners.[15,16] Critical evaluation of one's own knowledge is essential to target and make the best use of limited time.[16] Physicians who have been exposed to self-directed learning as students, tend to continue in that mode and better keep up to date in later practice.[17] Therefore, students should be actively involved in learning, rather than be passive recipients of information.[18] Current theories in the science of learning emphasize the importance of active learning. Learning is facilitated by allowing learning to take place in meaningful contexts, by activating prior knowledge, by requiring the student to engage actively in the learning process and by arousing intrinsic motivation.[19,20] These cognitive psychological theories are finding increasing support from research findings in the area of neurobiology.[21,22] An instructional method that has been developed to stimulate active self-directed lifelong learning is problem-based learning (PBL). Students play an active role in generating learning issues, deciding how they will study them, and evaluating what they have learned. This process supports the development of students' self-directed, lifelong learning skills.[23,24] The students are encouraged to take substantial responsibility for their own learning. Independent and active learning is stimulated by discussing problems in small groups. First, the students are stimulated to discuss a problem in the tutorial group (phase 1). A problem consists of a description of a set of phenomena needing some kind of explanation.[25]

However, during the discussion some questions remain unanswered. These questions serve as learning issues (phase 2). Phases 3 and 4 are devoted to the student's individual study. Students search the literature or other sources of information that seem relevant to the learning issues (phase 3). Subsequently, students study the information they have found and prepare a report (phase 4) to present their findings to the tutorial group during the reporting phase. Finally, in phase 5, when the group meets again, students check whether the results of their self-directed study have enabled them to understand and explain the problem (reporting phase). Ideally, this should result in students adopting deep learning processes leading to better structured and organized knowledge which will be better transferred and retrieved.[20,26-28] It would therefore be desirable to use an assessment method that tests long term retrieval and structure of knowledge in stead of pure memorization of knowledge.

Maastricht

The Maastricht medical school started in 1974 and adopted PBL as the instructional method.[29] The Maastricht medical curriculum offers a six-year programme. The students enter medical school directly after secondary education. The first four years of the programme consist of interdisciplinary units which usually last six weeks. For every unit a so-called block book is provided consisting of a number of problems or cases related to the content of the unit.[30] A group of 8 to 10 students, the so-called tutorial group, meets twice a week for a two-hour session guided by a tutor.[31] In one tutorial session the students will use a systematic procedure to analyse at least one problem.[25] In between tutorial group sessions students can take part in a longitudinal integrated skills training programme and/or attend lectures and laboratory sessions.[32,33] The remaining time (60%) can be used for (self-)study purposes.[34,35] Students may study individually or in groups and all available resources (books, articles, the internet, audio, video and (staff) experts) may be used to gather information. In the next tutorial session students present and discuss the material they have found and prepared.[20,31] Years 5 and 6, the clinical phase, consist of clinical clerkships. Students rotate through the major clinical disciplines, including family medicine.[36]

The need for progress testing

At the end of each unit an examination is administered which contains questions related to the content of the unit. The content of this test reflects, as closely as possible, the goals for the specific unit. Originally, the emphasis in the evaluation of student achievement was exclusively on these end-of-unit examinations. Soon it became clear that these tests had become the major driving force behind student learning. Students simply "studied for the test" in order to increase their chance of passing.[37] This invited students to peak from hurdle to hurdle, "wiping the hard disk" after one exam to prepare for the next test and in this way trying to maximize their chances of success.[2,38] As a result, the tests discouraged individual learning paths, reinforced rote memorization and detracted from the time students spent studying for other purposes than tests. In order to maintain the educational philosophy, the direct connection between a specific part of the educational programme and assessment had to be severed. To achieve this, the progress test (PT) was developed. Wijnen introduced the concept of progress testing in 1976 and since the 1977-1978

academic year progress tests have been a fixture in the Maastricht Faculty of Medicine's assessment programme.[29,39-42]

The progress test

The progress test (PT) is best characterized as a comprehensive final examination in medicine reflecting the cognitive final objectives of the curriculum. It samples knowledge across all disciplines and content areas relevant for the medical degree. As a result the test has no direct link with any specific course or unit. The PT is aimed at assessment of the development of students' medical knowledge (within the framework of the final objectives of undergraduate medical education). Four times per academic year (September, December, March and May) all students (years 1 through 6) take the same PT simultaneously. For each occasion a new test is constructed, consisting of about 250 items that a newly graduated MD should be able to answer correctly. Questions for the PT are written by faculty members from all departments of the medical school. They are of all taxonomic levels and may include facts and figures or problem vignettes. The knowledge tested should be functional for the student upon graduation. Questions that can only be answered when the answer was memorized the night before are of little value in a PT. All PT questions are critically reviewed on content, wording and relevance by a review committee.[43,44] Virtually all submitted items are reformulated and many are rejected before test administration. All items must be accompanied by a reference to the literature. To ensure that tests are equivalent, a blueprint is used, derived from the International Classification of Diseases.[45-47] The test blueprint determines the required number of items for each category.[46] The distribution of the items among the categories is based on morbidity data, the amount of space devoted to the various subjects in general medical textbooks, and a survey among the departments.[46,48] For practical reasons (large numbers of both items and students) test items are restricted to a closed format, the (multiple) True / False / Question Mark format.[49] Students are discouraged from blind guessing by the use of formula scoring to calculate the test result. The test result is calculated by subtracting the number of incorrect answers from the number of correct answers. Students can use a question mark option (zero points) to indicate that they do not (yet) know the answer.

In Maastricht, the PT is administered in an indoor tennis court complex on a Wednesday from 09.00 till 13.00 hours. Students can use the full 4 hours but on average they sit 2 hours to complete the PT. First-year students spend less time (1 hr 40 mins) and sixth-year students take more time to complete the PT (2 hr 50 mins). Food and drinks are allowed. After test administration, students take the booklets home. At 13.00 hours the answer key is available for the students. All students are invited to criticize items when they find conflicting evidence in the literature.[50,51] Students' comments and test statistics are again reviewed by the review committee. Flawed items are dropped after consultation with the department that submitted the item. Subsequently, the final test scores are calculated and the correct minus incorrect score is expressed on a percentage scale. Information on their performance is sent to the students and the education office. Every student receives his/her individual results accompanied by the mean results for the whole class. Total test score, mean ICD category scores and discipline scores are presented as well as the pass/fail

decision. Eliminated items and answer key changes are shown also. By comparing their own score to the mean class score, students can identify gaps in their knowledge. Mean ICD category score and discipline scores per class are also reported to departments and curriculum committees. Item scores are reported to the department of origin together with students' comments and an item report by the review committee.[47] Over the course of the six-year curriculum students sit 24 PTs. The test results reflect how far students have progressed towards the final objectives of undergraduate medical education, and thus yield valuable information to guide individual students and evaluate the curriculum.[31,42] The cross-sectional and longitudinal design of the PT gives it powerful research potential.

The concept of progress testing

Independently of the progress test, the Quarterly Profile Examination was developed at the University of Missouri Medical School in Kansas City.[52,53] More recently, the Personal Profile Index was introduced at McMaster University.[54,55] Progress testing is not confined to undergraduate medical education. The concept is very broadly applicable. Nowadays, longitudinal assessment methods are frequently used in various areas of undergraduate, postgraduate and continuing medical education.[56-60] Another area where longitudinal measurements are commonly used is quality of life research.[61,62] Probably the best known and oldest application of the concept is to be found in paediatrics. When a baby is born, its weight, height and head circumference are measured. The results are plotted on a growth chart and compared with the average reference group. Over the early years of life the child will be measured frequently and regularly. Whenever the results are not within an acceptable range, action will be taken. When this happens, the child will eventually be seen by a doctor to find out what is wrong. Clearly, the concept on which progress testing is based, i.e. that of repeated measurement to monitor growth and development, is not new and not limited to knowledge or education.

This dissertation will focus on the use of progress testing within the context of (undergraduate) medical education. In chapter two a model will be presented on the basis of which the utility of the concept will be studied.

References

1 Van der Vleuten CPM. Beyond intuition [Inaugural lecture Maastricht University]. Maastricht: Datawyse, 1996.

2 Van der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education* 1996; **1**: 41-67.

3 Epstein RM, Hundert EM. Defining and assessing professional competence. *JAMA* 2002; **287**: 226-35.

4 Van der Vleuten CPM, Scherpbier AJJA, Dolmans DHJM, Schuwirth LWT, Verwijnen GM, Wolfhagen HAP. Clerkship assessment assessed. *Medical Teacher* 2000; **22**: 592-600.

5 Swanson DB, Norcini JJ, Grosso LJ. Assessment of clinical competence: Written and computer-based simulations. *Assessment and Evaluation in Higher Education* 1987; **12**: 220-46.

6 Neufeld VR, Norman GR, Feightner JW, Barrows HS. Clinical problem-solving by medical students: A cross-sectional and longitudinal analysis. *Medical Education* 1981; **15**: 315-22.
7 Waldrop MM. The necessity of knowledge. *Science* 1984; **223**: 1279-82.
8 Norman GR, Tugwell P, Feightner JW, Muzzin LJ, Jacoby LL. Knowledge and clinical problem-solving. *Medical Education* 1985; **19**: 344-56.
9 Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise: Theory and implications. *Academic Medicine* 1990; **65**: 611-21.
10 Bordage G. Elaborated knowledge: A key to successful diagnostic thinking. *Academic Medicine* 1994; **69**: 883-5.
11 Regehr G, Norman GR. Issues in cognitive psychology: Implications for professional education. *Academic Medicine* 1996; **71**: 988-1001.
12 Van de Wiel MWJ. Knowledge encapsulation. Studies on the development of medical expertise [PhD dissertation Maastricht University]. Wageningen: Ponsen & Looijen, 1997.
13 Van der Vleuten CPM, Newble DI. How can we test clinical reasoning? *Lancet* 1995; **345**: 1032-4.
14 Cox K. Knowledge which cannot be used is useless. *Medical Teacher* 1987; **9**: 145-54.
15 Ludvigsson J. A curriculum should meet future demands. *Medical Education* 1999; **21**: 127-8.
16 Abrahamson S, Baron J, Elstein AS, Hammond WP, Holzman GB, Marlow B, Taggart MS, Schulkin J. Continuing medical education for life: Eight principles. *Academic Medicine* 1999; **74**: 1288-94.
17 Davis D, Thomson MA. Implications for undergraduate and graduate education derived from quantitative research in continuing medical education: Lessons learned from an automobile. *The Journal of Continuing Education in the Health Professions* 1996; **16**: 159-66.
18 Glaser R. The maturing of the relationship between the science of learning and cognition and educational practice. *Learning and Instruction* 1991; **1**: 129-44.
19 Van der Vleuten CPM, Dolmans DHJM, Scherpbier AJJA. The need for evidence in education. *Medical Teacher* 2000; **22**: 246-50.
20 Norman GR, Schmidt HG. The psychological basis of problem-based learning: A review of the evidence. *Academic Medicine* 1992; **67**: 557-65.
21 Sara SJ. Retrieval and reconsolidation: Toward a neurobiology of remembering. *Learning & Memory* 2000; **7**: 73-84.
22 Tronel S, Sara SJ. Mapping of olfactory memory circuits: Region-specific c-fos activation after odor-reward associative learning or after its retrieval. *Learning & Memory* 2002; **9**: 105-11.
23 Barrows HS, Tamblyn RM. *Problem -based learning. An approach to medical education.* New-York: Springer Publishing Company, 1980.
24 Barrows HS. A specific problem-based, self-directed learning method designed to teach medical problem-solving skills, and enhance knowledge retention and recall. In: Schmidt HG, Volder de ML, editors. *Tutorials in problem-based learning. A new direction in teaching the health professions.* Vol. 1. Maastricht/Assen: Van Gorcum; 1984: 16-32.
25 Schmidt HG. Problem-based learning: Rationale and description. *Medical Education* 1983; **17**: 11-6.
26 Norman GR, Schmidt HG. Effectiveness of problem-based learning curricula: Theory, practice and paper darts. *Medical Education* 2000; **34**: 721-8.
27 Albanese M. Problem-based learning: Why curricula are likely to show little effect on knowledge and clinical skills. *Medical Education* 2000; **34**: 729-38.
28 Colliver JA. Effectiveness of problem-based learning curricula: Research and theory. *Academic Medicine* 2000; **75**: 259-66.

29 Knegtmans PJ. *De medische faculteit Maastricht: Een nieuwe universiteit in een herstructureringsgebied, 1969-1984. [The Maastricht medical school: A new university in a restructurisation area, 1969-1984].* Assen: Van Gorcum, 1992.

30 Dolmans DHJM, Snellen-Balendong HAM, Wolfhagen HAP, Van der Vleuten CPM. Seven principles of effective case design for a problem-based curriculum. *Medical Teacher* 1997; **19**: 185-9.

31 Van der Vleuten CPM, Scherpbier AJJA, Wijnen WHFW, Snellen HAM. Flexibility in learning: A case report on problem-based learning. *International Higher Education* 1996; **2**: 17-24.

32 Van Luijk SJ. Al doende leert men. [Practice makes perfect] [PhD dissertation Rijksuniversiteit Limburg]. Maastricht, 1994.

33 Scherpbier AJJA. Kwaliteit van vaardigheidsonderwijs gemeten [Assessing the quality of skills training] [PhD dissertation Maastricht University]. Maastricht: Datawyse, 1997.

34 Van den Hurk MM. Individual study in problem-based learning [PhD dissertation Maastricht University]. Maastricht: Datawyse, 1999.

35 Gijselaers WH, Schmidt HG. Effects of quantity of instruction on time spent on learning and achievement. *Educational Research and Evaluation* 1995; **1**: 183-201.

36 Wolfhagen HAP. Kwaliteit van klinisch onderwijs [Quality of clinical education] [PhD dissertation Maastricht University]. Maastricht: Datawyse, 1993.

37 Van Berkel HJM, Nuy HJP, Geerlings T. The influence of progress test and block tests on study behaviour. *Instructional Science* 1995; **22**: 317-33.

38 Leeder SR, Feletti GI, Engel CE. Assessment - help or hurdle? *Programmed Learning & Educational Technology* 1979; **16**: 308-15.

39 Wijnen W. Einddoel-toetsen: Waarom en hoe? [Assessing final objectives: Why and how?]. *Onderzoek van Onderwijs* 1977; **6**: 16-9.

40 Wijnen WHFW. Onder of boven de maat. Een methode voor het bepalen van de grens voldoende/onvoldoende bij studenten. [Above or below par. A method to determine the pass/fail cutoff point in student assessments] [PhD dissertation Rijksuniversiteit Groningen]. Amsterdam: Swets & Zeitlinger, 1971.

41 Wijnen WHFW, Van der Vleuten CPM. Toetsing: Hordenloop of voortgangskontrole? [Assessment: Hurdle-race or progress control?]. *Universiteit en Hogeschool* 1985; **31**: 270-9.

42 Van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher* 1996; **18**: 103-9.

43 Van Hessen PAW, Verwijnen GM. Necessity of a test review committee in test construction. *Paper presented at: The international symposium on evaluation in medical education.* Beer Sheva, Israel; 1987.

44 Van Hessen PAW, Verwijnen GM. Toetsen getoetst. Het beoordelen van toetsvragen in Maastricht [Assessment of tests. The judgement of test-items in Maastricht]. *Bulletin Medisch Onderwijs* 1989; **8**: 100-5.

45 World Health Organization. *The international classification of diseases.* Vol. 1. 9th ed. Michigan: Edwards Brothers Inc., 1978.

46 Verwijnen M, Imbos T, Snellen H, Stalenhoef B, Pollemans M, Van Luyk S, Sprooten M, Van Leeuwen Y, Van der Vleuten C. The evaluation system at the medical school of Maastricht. *Assessment and Evaluation in Higher Education* 1982; **7**: 225-44.

47 Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Schuwirth LWT, Van der Vleuten CPM. Quality assurance in test construction: The approach of a multidisciplinary central test committee. *Education for Health* 1998; **12**: 49-60.

48 Verwijnen GM, Fröberg-Gresnich C. *Jaarverslag over het academisch jaar 1984/1985 van de voortgangstoets-beoordelingscommissie [Annual report of the progress test review committee 1984/1985].* [PES-publ.nr. 86-117] Maastricht: Universiteit Maastricht, Faculteit der Geneeskunde; 1986.
49 Ebel RL, Frisbie DA. *Essentials of educational measurement.* 5th ed. New Jersey: Englewood Cliffs, 1991.
50 Prince CJAH, Visser K. The student as quality controller. In: Scherpbier AJJA, Van der Vleuten CPM, Rethans JJ, Van der Steeg AFW, editors. *Advances in Medical Education.* Dordrecht / Boston / London; 1997: 15-8.
51 Verwijnen GM. De student als kwaliteitsbewaker. De rol van de student bij de kwaliteitsbewaking van de Maastrichtse voortgangstoets [The student as quality controller. The student's role in quality assurance of the Maastricht progress test]. *Bulletin Medisch Onderwijs* 1994; **13**: 87-95.
52 Willoughby TL, Dimond EG, Smull NW. Correlation of quarterly profile examination and national board of medical examiner scores. *Educational and Psychological Measurement* 1977; **37**: 445-9.
53 Arnold L, Willoughby TL. The quarterly profile examination. *Academic Medicine* 1990; **65**: 515-6.
54 Blake JM, Norman GR, Keane DR, Mueller CB, Cunnington J, Didyk N. Introducing progress testing in McMaster university's problem-based medical curriculum: Psychometric properties and effect on learning. *Academic Medicine* 1996; **71**: 1002-7.
55 Blake JM, Norman GR, Kinsey E, Smith M. Report card from McMaster: Student evaluation at a problem-based medical school. *Lancet* 1995; **345**: 899-902.
56 Van Leeuwen YD. Growth in knowledge of trainees in general practice. Figures on facts [PhD dissertation Maastricht University]. Maastricht: Datawyse, 1995.
57 Ram P. Comprehensive assessment of general practitioners [PhD dissertation Maastricht University]. Maastricht: Unigraphic, 1998.
58 Pollemans M. Kennistoetsing bij huisartsen [Testing knowledge of general practitioners] [PhD dissertation Maastricht University]. Maastricht: Datawyse, 1994.
59 Schuwirth LWT, Schrander JPP, Van der Vleuten CPM. Assistententoets kindergeneeskunde: Een beschrijving van de psychometrische eigenschappen. [The national examinaton for residents in pediatrics: A description of its psychometric properties]. *Bulletin Medisch Onderwijs* 1993; **12**: 146-51.
60 Shen L. Progress testing for postgraduate medical education: A four-year experiment of american college of ostepathic surgeons resident examinations. *Advances in Health Sciences Education* 2000; **5**: 117-29.
61 Kiebert GM, Curran D, Aaronson NK. Quality of life as an endpoint in EORTC clinical trials. *Statistics in Medicine* 1998; **17**: 561-9.
62 Bottomley A, Therasse P. Quality of life in patients undergoing systemic therapy for advanced breast cancer. *Lancet Oncology* 2002; **3**: 620-8.

Chapter 2

Research questions

A model for determining test utility

The purpose of most examinations is to determine the competency of the students. Examinations should guarantee that competent students pass and incompetent students fail. As stated earlier, competence is content specific. What is being measured by an individual item in a test depends on the cognitive processes that are triggered in the person trying to answer the question. These processes will not only vary from item to item but also from person to person. What may be the result of simple recognition for one person, may be the result of a reasoning process for another.[1] Therefore, assessment of medical competence is difficult and perfect assessment an illusion. No single assessment method offers a panacea.[2] There is no consensus on what is the "best" method for assessing competence. As a result, countless instruments have been developed to assess competence, varying from multiple choice questions to observation of performance in real practice. Each assessment method is a compromise between what is desirable and what is achievable. Selection of assessment methods should depend on the skills to be assessed. In most cases, a blend of methods will be desirable.[3-5]

Assessment serves personal, institutional, and societal goals. It is used to 1) pass or fail a student, 2) grade or rank a student, 3) provide feedback to students, teachers, curriculum and institution, and, 4) motivate students, teachers and institution.[6,7] Distinctions between the different goals of assessment are often blurred in educational and testing practice but nevertheless very important. Depending on the goal of the test, the what, how, why, when and who of assessment differ.[6] Not every form of assessment is useful for all of the above mentioned goals. In other words, the utility of a test depends on its purpose and the properties of the assessment method. Van der Vleuten has developed a conceptual model to define the utility of an assessment method.[1] This model relies on five aspects: 1) reliability, 2) validity, 3) educational impact, 4) acceptability, and 5) costs. The utility of an assessment method depends strongly on all five aspects, as is reflected by the formula:

$$U = R \times V \times E \times A \times C$$

Perfect utility is unattainable, but for an assessment method to have any utility, none of the variables can equal zero. In practice, compromises will have to be made and depending on the context and purpose of the assessment not all elements will be of equal importance. In different situations, the variables will be associated with differential weights:

$$U = R_{w_r} \times V_{w_v} \times E_{w_e} \times A_{w_a} \times C_{w_c}$$

For a high stakes examination most weight will be put on reliability, whereas for a final decision based on many assessments in an in-training evaluation programme, one may be prepared to compromise on reliability in favour of, for example, educational impact. Of

course, this formula is only intended as a conceptual model. It is not to be used as an actual algorithm, because most of the elements cannot be quantified. However, the formula can clarify the trade-offs involved in the implementation of assessment methods in a medical curriculum.
In the following section this model will be used to evaluate the utility of the progress test. Two of the elements (validity and educational impact) have resulted in seven research questions which are studied and discussed in the following chapters.

Research questions

Reliability
Reliability refers to the precision of a measurement or, in the case of assessment, the reproducibility of test results.[8,9] The purpose of any assessment is to draw conclusions about the ability of the candidate beyond the particular sample of items and test conditions. The candidate's test score should be stable and reproducible across different but similar samples of items, raters, testing sites, time of day, patients, et cetera.[10] Reproducibility can be impaired due to various sources of error. Many studies on reliability have been published and the general conclusion is that the main source of interference in all test formats is the fact that competence measurement is content-specific. In some cases, this may not be very surprising when content areas are very different (for example biochemistry and surgery), but the problem also occurs with specific content areas (for example inguinal hernias and epigastric hernias). Knowing how a candidate has handled a patient's problem or answered a question is not a good predictor of how the candidate will deal with another problem or question, even if they are related to the first question. The content specificity of medical competence implies that tests consisting of a small sample of items will be unreliable. Since the progress test samples the complete domain of factual knowledge of a curriculum, many items are needed. Wide sampling of content across the area of interest is imperative to obtain stable and reproducible scores. Several studies have been conducted to establish the reliability of progress testing methods. In general, the reported reliabilities were high. Cronbach's alphas varying between 0.70 to 0.80 were achieved within classes and across classes they were usually above 0.95.[11-15] No specific research question about the reliability of progress tests was formulated for this dissertation.

Validity
Validity refers to "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests".[16] Cronbach formulated validity as "the accuracy of a prediction or inference made from a test source".[17] To put it more simply: a valid test measures what it is intended to measure.[9] For a test to be valid, it should be reliable.[18] Validity is a property of the way test scores are interpreted. It is not a property of the test itself. This means that a test can have many different validities, depending on how the scores are used.[19] Traditionally, three main categories are distinguished: content validity, construct validity and criterion validity.

Content validity
Ebel stressed the importance of content validity ("intrinsic rational validity").[18] He argued that an examination should consist of problems that candidates encounter during training and after certification. The choice of test material should be rational, clear and open to discussion. The progress test is aimed at the assessment of any increase in students' medical knowledge (within the framework of the final objectives of undergraduate medical education). The progress test should contain questions that a newly graduated MD should be able to answer correctly and the knowledge tested should be functional for the student upon graduation. Questions which can only be answered when the answer was memorized the night before are of little value. All departments of the Maastricht medical school devise progress test items which are entered into a central item bank without being evaluated. Some eight months before the date of the progress test, some 400-500 items are drawn and peer reviewed.

Research question 1
What measures are taken to assure the content validity of the progress test?

In *chapter three* the pre- and post-test review processes as well as the construction of the progress test are described in detail.

Until 1994, no criteria were available that the progress test review committee could apply to determine the relevance of an item. In "Blueprint 1994, training of doctors in the Netherlands", a national document, the final objectives of undergraduate medical education in The Netherlands have been stated.[20]

Research question 2
Does the progress test reflect the final (cognitive) objectives of undergraduate medical education?

The study reported in *chapter four* compared the content of the PT to the objectives of the Blueprint.

Construct validity
Construct validity is concerned with the validity of a hypothetical construct as measured by a test.[21] One of the goals of medical education is to gain medical knowledge. As students progress through the curriculum, their medical knowledge should increase. The concept of progress testing is that progress tests monitor the growth of medical knowledge over the course of the medical curriculum. If the progress test does indeed measure medical knowledge, the average test scores should rise (on successive tests) with additional years of training. A similar increase should be seen when the scores of different cohorts of students on the same test are compared, i.e. first year students should score considerably lower than students in their sixth year of training. Only some of the published studies on the PT

mention the construct of growth. Shen demonstrated construct validity for postgraduate progress testing by evaluating the psychometric properties of the test results using a Rasch model.[14] Van Leeuwen and Pollemans performed a study using cross-sectional data and found support for the construct validity of the knowledge test for general practice.[12,13] For the PT used in undergraduate medical education it has been proven that differences between levels of expertise measured by the test were not attributable to students becoming more mature or general individual development over the years. Progress test scores were found to reflect a level of specific medical knowledge.[22] Imbos showed that the progress test discriminated between students of different levels of ability, albeit that his research and conclusions were mainly focused on the first year of medical education.[11] Furthermore, based on the progress test a progress model (IRT) has been developed by Albers et al. to describe successive test results on longitudinal assessment methods and predict individual and group performance.[23] The progress test was found to be suitable for comparing growth of knowledge of different groups of students.[24,25] The results of these studies provide evidence for the construct validity of the progress test.

Research question 3
Is the progress test capable of measuring growth of medical knowledge over successive years of training?

Chapter five deals with the growth of medical knowledge during six years of undergraduate medical education. Overall growth and growth on sub-scores (clusters) was shown. The results are related to the content of the medical curriculum in Maastricht.

Although progress testing is widely used in undergraduate medical curricula all over the world, its concept is not confined to undergraduate medical education. In the field of medical education, forms of progress testing are used in several post-graduate educational programmes[12-14,26] and in continuing medical education[27]. The concept is also used in pharmacy[28], psychology, dentistry and veterinary medicine curricula. However, all applications of the concept are in the domain of cognitive (factual) knowledge testing. A strong argument in favour of the construct validity of PTs would be evidence that the concept of progress testing can also be applied in other domains such as the assessment of clinical skills.

Research question 4
Can progress testing be applied in assessing knowledge of skills?

One of the most valid and best known methods for assessment of clinical skills is the Objective Structured Clinical Examination (OSCE).[29] It is used in undergraduate and postgraduate medical education.[30-32] Like all forms of assessment the OSCE shows a large variability in performance between stations, mainly due to the content specificity of competence. Many stations are needed to reach reliable test results due to the large

variability in performance between cases and tasks. The OSCE has an important drawback compared to paper-and-pencil tests because the organisation is more complex, more expensive, and more hours of testing time are needed.[31,33,34] These drawbacks are a barrier to large-scale application of the OSCE on logistical and financial grounds.

Miller has suggested that knowledge of skills ("knows how") should be present before a skill can be performed properly ("shows how").[3] A technically correct performance of a procedure requires procedural and technical knowledge. In addition, one must know the indications for the procedure and be able to interpret the findings. Several studies investigated the correlation between "knows how" and "shows how" by comparing test results of OSCEs and knowledge tests of skills (KTS) in order to find more efficient alternatives for the OSCE. Reasonably high true correlations between written KTSs and OSCEs were reported in various studies as well as good correlation between a KTS and actual medical performance evaluated through video observation.[35-41]

Chapter six illustrates the way progress test questions can be used to assess knowledge of clinical skills as a first step towards longitudinal testing of knowledge of skills.

Criterion validity
The progress test is used in a formative way to inform students about their progress, possible gaps in their knowledge and their relative position with regard to the end-objectives. The test is also used to reach pass/fail decisions for individual students. The validity of these decisions is dependent on standards. Standard setting requires judgement somewhere along the process.[42,43] The choice of a passing score is always arbitrary. In the end, no compelling reasons or evidence can be given why the passing score should not be set a little higher or a little lower.[44] Examinees just below the passing score do not differ substantially from examinees just above the passing score. Although it is not possible to establish unequivocally the validity of a standard (establishing its correctness), evidence on a standard's credibility and defensibility can be collected and used to support the use of the standard for a particular purpose.[45]

Research question 5
How reliable and credible is a content-based standard for the progress test?

Although the quality control process used in Maastricht is very thorough and test construction is highly standardized, successive PTs vary in degree of difficulty. This hampers the use of absolute standards. In order to minimize the number of false positive and false negative decisions, a normative approach, with relative cut-off scores being defined by the overall performance of a class, was used for 20 years.[46,47] It was essentially based upon two assumptions: 1) it is essential to involve the quality of the measurement (the test) in the standard setting procedure, and 2) the average candidate is assumed to be able to pass the test. This resulted in a passing score that was based on the standard error of measurement (SEM) of the test concerned (mean minus 1.96 times SEM).[46] For the

progress tests, which have an average reliability of 0.75, the mean score minus one standard deviation has always been used.[47,48] The latter procedure was easier to explain and more transparent for teachers and students. The distribution of progress test scores is approximately Gaussian resulting in a more or less fixed percentage of students who fail the test (roughly 16%) for every progress test.[45,47] The advantage of such an approach is that variations in test difficulty are automatically corrected for and examinees are not left to face the music of unreliable tests. This approach also has several disadvantages: 1) the passing score is not known in advance; 2) a number of students fail, regardless of their abilities; 3) examinees can deliberately influence the passing score; and 4) heterogeneity of the student population reduces the validity of the standard.[45,47] These disadvantages of the relative standard were considered to be of major importance, although it is a matter of opinion whether "not knowing the passing score in advance" is really a disadvantage and in practice students rarely "deliberately influence a relative passing score", which would be easy to detect anyway. Eventually, a PT re-examination was introduced. Also due to political pressure (especially from students[49]) it became inevitable to introduce an absolute passing score for the PT. Research into the possible effects of a fixed absolute standard for progress tests had shown that major fluctuations in failure rates (ranging from 2.5% to 50%) would accompany such a standard setting procedure.[47] Nevertheless, a fixed standard was calculated on the basis of past test performance during eight years[47] and introduced in September 1996.

Because of the comprehensive nature of progress tests and because each test administration requires a different passing score for each class, setting an absolute passing score is quite complicated. A possible alternative would be a content-based standard setting procedure based on item judgement by a panel of experts, such as an Angoff procedure. Two Angoff procedures with different panels judging the same progress test items were performed. The two studies that investigated these alternatives are described in *chapters seven and eight.*

Educational impact; influence on learning

For students, academic success is defined by the examinations.[50] What, how, when and how much students learn is to a large extent directed and shaped by tests.[51-55] This "law" describes one of the strongest relationships in education.[9] It is especially true if test results have far-reaching consequences for students' progress and thus their future. Students will study whatever they are tested on and are unlikely to study what they are not tested on; learning is primarily test-driven. Students will learn in the way that appears to yield the best results with maximum efficiency; usually with the least effort to be able to cope with competing interests. Any strategy that appears to offer the best chances of success is the one that is most likely to be used. This means that students tend to decide to memorize facts to know everything about one specific subject because it is the examiner's favourite topic or to skip all educational activities to be able to study. For students the assessment programme is the curriculum. [50,56] Assessment of student achievement should be congruent with sound educational principles. If the educational principles of the curriculum are not reflected in and reinforced by the assessment programme, the "hidden curriculum" of

assessment objectives will prevail.[57,58] Assessment drives learning not only through its content and format, but also through the information that is given to students afterwards and the way the different tests are programmed.

Assessment drives learning through its content

What is tested is considered essential for students to learn at the cost of matters that are not tested.[52,59,60] If a demonstration of practical skills is examined, students will practice these skills. If knowledge of facts and figures is tested, that is what they will learn and memory reproduction will be trained. Questions that focus on the application of knowledge, ask for management decisions or the probability of a diagnosis will drive students to prepare differently and focus on concepts instead of memorization of facts.[61] You will get out of the test what you put into it.[1]

Assessment drives learning through its format

The literature is rich with publications arguing that the method of measuring determines what is being measured. MCQs are thought to have limited capacity to assess higher order cognitive skills and OEQ or oral examinations are considered to be preferable. The empirical evidence does not support these intuitive beliefs and McGuire summarizes these findings very firmly: "The format of a question and the technique it employs have virtually nothing to do with what it measures."[61,62] However, the format of an examination can have a tremendous effect on learning.[51] A performance based test like the OSCE should stimulate students to practice and master clinical (practical) skills.[6,29] Van Luijk showed that the manner in which the OSCE was organized and the observations scored by the examiners had a major influence on how students prepared for the OSCE.[63,64] Although students obtained good results, it was obvious that their understanding of the skills they were demonstrating was superficial. For example, students told the examiner what they were supposed to hear, while not having the stethoscope in their ears or listening at the wrong places on the body. The OSCE stations tested skills in isolation (i.e. auscultation of the heart in one station, interviewing skills in another) and students appeared to memorize the checklists used by examiners in previous years. These checklists focused on very detailed procedural knowledge which could be learned by heart. The tactic used in preparing for such a test is similar to the well known "cramming", which is associated with many written examinations. This way, an OSCE stimulates memory reproduction instead of the learning of skills by practice. In response to these findings, the format of the OSCE was changed; stations were integrated and less detailed checklists were developed.[64]

Assessment drives learning through feedback

Test results are most frequently used and expressed as a decision tool. Passing or failing a test is only one, but unfortunately fairly overestimated, aspect of assessment. This tends to detract from the educational evaluative value of test results. Students remain ignorant of their strengths and weaknesses if they are only told that they have passed or failed a test and areas that require extra work or remedial teaching are not identified. Students are unable to learn from their mistakes and will remember their wrong answer as being right or at best as "not exactly sure". The information given to students should include the test items,

literature references and answer keys. Only then will students be able to correct their false assumptions about certain topics or signal problems with test items, conflicting literature or course material. The feedback value of tests is also important for teachers and curriculum designers. Test results indicate to what extent students have achieved the educational goals of a teaching programme. Items that are answered incorrectly by most of the students may indicate that a particular topic is addressed inadequately during the course or not covered by textbooks or other learning materials.[65-68]

Assessment drives learning through its programming

Miller described the introduction of a weekly quiz on Friday by a Department of Pathology at a medical school in the USA. Soon the Department of Pharmacology noticed a decrease in student attendance at the scheduled Thursday afternoon laboratory session. The teachers discovered that many students used Thursday afternoon for preparing the Pathology quiz and reacted with the administration of unannounced laboratory examinations to restore earlier attendance rates.[55] This way tests were used as instruments to get students' time and attention. Test programming is a powerful means to force students to study regularly or attend certain classes. Too many examinations within a limited period of time will cause competition between the tests. The test most likely to influence the academic success of a student will get most attention. An examination period of a few months once a year will result in peak load and force students to make choices. Students appear to start their preparation activities about 3 to 4 weeks before a test. This results in competition between tests and usually the last planned exam(s) lose(s).[53] In the end all this could result in partially achieved educational goals and delayed study progress. Although repetition of examinations may seem fair to students, it encourages minimal learning strategies and "scouting" at first attempts with minimal preparation. For there is always a chance of success and a new chance.[1,53] Furthermore, the student promotion regulations directly influence student behaviour because they define academic success. Students make use of compensatory rules across assessment and know what can be skipped or has first priority. They will respond immediately and strategically to changes in the regulations but the educational effects are only partially predictable. Exams should be spread over the year and not compete for students attention. Failing one test should not be disastrous for a student. Compensatory rules should exist across examinations and assessment methods and one opportunity per year to repeat an examination will cause the least delay in study progress.[53]

The described phenomena can be used strategically to reinforce desirable learning behaviour by applying "measurement-driven instruction".[69] It is a powerful method albeit with unpredictable educational consequences. Another way of trying to prevent students from following a "hidden curriculum" of undesirable assessment-based objectives are longitudinal final objective assessment methods such as the progress test, quarterly profile examination and the personal profile index. These tests are designed to have no or minimal influence on study patterns. Several studies have confirmed that the influence of these assessment methods on study behaviour is minimal.[70-74] However, these studies are not fully applicable to the Maastricht PT and left some questions unanswered. The design and introduction of the progress test was triggered by the need to maintain the educational

philosophy of PBL in medical education. The desired learning behaviours (active self-directed and open-discovery learning) of students should not be hampered by the progress test. It is of vital importance to find out if the PT has lived up to these expectations..

Research question 6
Does the progress test have a positive influence on learning behaviour?

In *chapter nine* some aspects of the educational impact of the progress test are described and discussed in relation to the impact of other tests that are used in the assessment of students.

Educational impact; influence on teaching

After the validity of a measuring instrument has been established, the instrument can be used as a source of information. A test that was first the object of research, can be turned into a research tool. Progress test results can be used by departments, teachers and curriculum committees to evaluate the educational programme. Test results of one class indicate which aspects are understood or known. Cross sectional data on a particular item shows differences between classes. This information indicates at what moment in the curriculum a topic is dealt with by students (which may not the moment curriculum designers had in mind when designing the curriculum) think it should be covered) and the influence of time on the retention of knowledge. It is also possible to evaluate the effects of specific courses or experiments in the curriculum. Pre-tests and post-tests are readily available as are students and items that can serve as controls.[15,75] Progress test items can be categorized in multiple ways to provide relevant information for a range of purposes. The combination of cross-sectional and longitudinal attributes in the design of the progress test provides powerful research potential. The progress test is not geared to a specific type of curriculum and therefore can be used to compare different curricula. Several comparisons at macro level were conducted in the past 25 years.[76-78]

Research question 7
Can the concept of progress testing be used to evaluate and compare educational programmes?

Chapter ten presents a detailed comparison between the progress tests results at two Dutch medical schools to shed more light on the possibilities and impossibilities of the instrument. In *chapter eleven* an international comparison of three progress tests is presented with correction for cultural bias and translation problems.

Acceptability
Irrespective of the quality of the assessment method, the acceptation by potential users will determine its utility. Teachers as well as students should accept, or rather appreciate, the

test results. This can only be achieved if they believe in the value of the test. Unless the value of the test is accepted, the test and its results will not be taken seriously and all efforts will be in vain. Without acceptance, an assessment method will ultimately not survive.[50,56,79] Examiners usually dislike highly structured assessments because it restricts their professional freedom and does not use their expertise. Furthermore, many teachers are anxious to have a say in what is assessed and how it is assessed. When test material is developed by others its acceptability to other teachers might be threatened.[80] On the other hand the time that is allocated for teaching and the time teachers spend on test development and scoring influence their feelings about (efficient) test approaches also.[1,5] For students, an important concept with regard to acceptability is "fairness". This concept appears to be tightly linked with the educational ideas about "validity". A fair assessment from a student's point of view is a valid measurement of what he or she judges to be meaningful and worthwhile to learn.[81] Acceptability for students is also strongly influenced by the purpose served by the test and the way the test results are used.[5] Students tend to see assessment exclusively as a summative tool which follows after learning and is used for pass/fail decisions only. This view is incompatible with the use of test results to correct misconceptions and enhance learning behaviour.

The progress test is used to make pass/fail decisions and supplies students with extensive feedback about their progress. This combination of functions and the relative importance of these functions may influence how acceptable the test is to students. The exact position and weight of the progress test results in pass/fail decisions has changed constantly due to modifications of local examination rules over the years. These factors and new national legislation on student grants appear to have been a major influence on the results of several surveys conducted since the introduction of the progress test in 1976.[82] In this dissertation the acceptability of progress testing is not addressed separately.

Costs

Assessment is costly and good assessment is definitely more costly. Test construction with built-in pre-test review and post-test quality control processes, development of simulations, training of (standardized) patients and examiners, test administration, data processing, feedback to students, staff and faculty and monitoring of effects are all time consuming and expensive activities. Costs can play an important role in choices between assessment methods. In fact, costs are one of the main reasons why compromises in assessment are, and have to be, made. In debates about the costs of assessment it is often forgotten that good assessment will drive and facilitate good learning. Assessment does not serve a different purpose than teaching and learning do. An investment in testing is worth the money and effort because, in the end, it will pay off.[1] In this dissertation no research question was included that addressed the cost of assessment.

In the following chapters nine studies are described which will provide answers to the seven research questions and provide insight into the different aspects of the utility of progress testing as a concept.

References

1 Van der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education* 1996; **1**: 41-67.
2 Case SM. Assessment truths that we hold as self-evident and their implications. In: Scherpbier AJJA, Van der Vleuten CPM, Rethans JJ, Van der Steeg AFW, editors. *Advances in Medical Education.* Dordrecht / Boston / London; 1997: 2-6.
3 Miller GE. The assessment of clinical skills / competence / performance. *Academic Medicine* 1990; **65**: S63-7.
4 Swanson DB, Norman GR, Linn RL. Performance-based assessment: Lessons from the health professions. *Educational Researcher* 1995; **24**: 5-11.
5 Norman GR, Van der Vleuten CP, De Graaff E. Pitfalls in the pursuit of objectivity: Issues of validity, efficiency and acceptability. *Medical Education* 1991; **25**: 119-26.
6 Harden RM. Assess students: An overview. *Medical Teacher* 1979; **1**: 65-9.
7 Epstein RM, Hundert EM. Defining and assessing professional competence. *JAMA* 2002; **287**: 226-35.
8 Nunnally JC. Theory of measurement error. *Psychometric theory.* 2nd ed. New York: Mc Graw-Hill Book Company; 1978: 190-224.
9 Van der Vleuten C. Validity of final examinations in undergraduate medical training. *British Medical Journal* 2000; **321**: 1217-9.
10 Van der Vleuten CP, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: Issues of reliability. *Medical Education* 1991; **25**: 110-8.
11 Imbos T. Het gebruik van einddoeltoetsen bij aanvang van de studie [Using assessment of final objectives at the start of a study] [PhD dissertation Rijksuniversiteit Limburg]. Maastricht: 1989.
12 Pollemans M. Kennistoetsing bij huisartsen [Testing knowledge of general practitioners] [PhD dissertation Maastricht University]. Maastricht: Datawyse, 1994.
13 Van Leeuwen YD. Growth in knowledge of trainees in general practice. Figures on facts [PhD dissertation Maastricht University]. Maastricht: Datawyse, 1995.
14 Shen L. Progress testing for postgraduate medical education: A four-year experiment of american college of ostepathic surgeons resident examinations. *Advances in Health Sciences Education* 2000; **5**: 117-29.
15 Van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher* 1996; **18**: 103-9.
16 American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for educational and psychological testing.* Washinton: AERA, 1999.
17 Cronbach LJ. Test validation. In: Thorndike RL, editor. *Educational measurement.* 2nd ed. Washington, D.C.: American Council on Education; 1971: 443-507.
18 Ebel RL. The practical validation of tests of ability. *Educational Measurement: Issues and Practice* 1983; **2**: 7-10.
19 Swanson DB, Norcini JJ, Grosso LJ. Assessment of clinical competence: Written and computer-based simulations. *Assessment and Evaluation in Higher Education* 1987; **12**: 220-46.
20 Metz JCM, Stoelinga GBA, Pels Rijcken - van Erp Taalman Kip EH, Van den Brand - Valkenburg BWM. *Blueprint 1994: Training of doctors in the netherlands. Objectives of undergraduate medical education.* Nijmegen: University Publication Office, 1994.
21 Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychological Bulletin* 1955; **52**: 281-302.

22 Fokkema F. *Het gebruik van voortgangstoetsen bij het vergelijken van medische curricula: Een begripsvaliditeitsstudie [The use of progress testing in comparing medical curricula: A study of construct validity].* [Heijmans Bulletin (HB-86-812-SW)] Groningen: Rijksuniversiteit Groningen; 1986.

23 Albers W, Does RJMM, Imbos T, Janssen MPE. A stochastic growth model applied to repeated tets of academic knowledge. *Psychometrika* 1989; **54**: 451-66.

24 Tan ES, Imbos T, Does RJMM. A distribution-free approach for comparing growth of knowledge. *Journal of Educational Measurement* 1994; **31**: 51-65.

25 Tan ES, Imbos T, Does RJMM, Theunissen M. An optimal, unbiased classification rule for mastery testing based on longitudinal data. *Educational and Psychological Measurement* 1995; **55**: 595-612.

26 Schuwirth LWT, Schrander JPP, Van der Vleuten CPM. Assistententoets kindergeneeskunde: Een beschrijving van de psychometrische eigenschappen [The national examinaton for residents in pediatrics: A description of its psychometric properties]. *Bulletin Medisch Onderwijs* 1993; **12**: 146-51.

27 Ram P. Comprehensive assessment of general practitioners [PhD dissertation Maastricht University]. Maastricht: Unigraphic, 1998.

28 Fourie S, Summers B, Löwes MMJ, Summers RS. Development of a student progress test for the BPharm offered by MEDUNSA, in partnership with TP. *Pharmaciae* 2002; **10**: 20-1.

29 Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education* 1979; **13**: 41-54.

30 Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, Brown M. Objective structured assessment of technical skill (OSATS) for surgical residents. *British Journal of Surgery* 1997; **84**: 273-8.

31 Van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine* 1990; **2**: 58-76.

32 Stillman P, Swanson D, Regan MB, Philbin MM, Nelson V, Ebert T, Ley B, Parrino T, Shorey J, Stillman A, Alpert E, Caslowitz J, Clive D, Florek J, Hamolsky M, Hatem C, Kizirian J, Kopelman R, Levenson D, Levinson G, McCue J, Pohl H, Schiffman F, Schwartz J, Thane M, Wolf M. Assessment of clinical skills of residents utilizing standardized patients. A follow-up study and recommendations for application. *Annals of Internal Medicine* 1991; **114**: 393-401.

33 Reznick RK, Smee S, Baumber JS, Cohen R, Rothman A, Blackmore D, Berard M. Guidelines for estimating the real cost of an objective structured clinical examination. *Academic Medicine* 1993; **68**: 513-7.

34 Carpenter JL. Cost analysis of objective structured clinical examinations. *Academic Medicine* 1995; **70**: 828-33.

35 Jansen JJ, Tan LH, Van der Vleuten CP, Van Luijk SJ, Rethans JJ, Grol RP. Assessment of competence in technical clinical skills of general practitioners. *Medical Education* 1995; **29**: 247-53.

36 Van der Vleuten CPM, Van Luyk SJ, Beckers HJM. A written test as an alternative to performance testing. *Medical Education* 1989; **23**: 97-107.

37 Scherpbier AJJA, Verwijnen GM, Schaper N, Dunselman GAJ, Van der Vleuten CPM. Vaardigheidsonderwijs nu en in de toekomst [Current and future skills training]. *Tijdschrift voor Medisch Onderwijs* 2000; **19**: 6-15.

38 Remmen R, Scherpbier A, Denekens J, Derese A, Hermann I, Hoogenboom R, van Der Vleuten C, van Royen P, Bossaert L. Correlation of a written test of skills and a performance based test: A study in two traditional medical schools. *Medical Teacher* 2001; **23**: 29-32.

39 Kramer AW, Jansen JJ, Zuithoff P, Dusman H, Tan LH, Grol RP, van der Vleuten CP. Predictive validity of a written knowledge test of skills for an OSCE in postgraduate training for general practice. 2002; **36**: 812-9.

40 Ram P, Grol R, Rethans JJ, Schouten B, van der Vleuten C, Kester A. Assessment of general practitioners by video observation of communicative and medical performance in daily practice: Issues of validity, reliability and feasibility. *Medical Education* 1999; **33**: 447-54.

41 Ram P, van der Vleuten C, Rethans JJ, Schouten B, Hobma S, Grol R. Assessment in general practice: The predictive value of written- knowledge tests and a multiple-station examination for actual medical performance in daily practice. *Medical Education* 1999; **33**: 197-203.

42 Livingston SA, Zieky MJ. A comparative study of standard-setting methods. *Applied Measurement in Education* 1989; **2**: 121-41.

43 Norcini JJ. Research on standards for professional licensure and certification examinations. *Evaluation and the Health Professions* 1994; **17**: 160-77.

44 Kane M. Validating the performance standards associated with passing scores. *Review of Educational Research* 1994; **64**: 425-61.

45 Norcini JJ, Shea JA. The credibility and comparability of standards. *Applied Measurement in Education* 1997; **10**: 39-59.

46 Wijnen WHFW. Onder of boven de maat. Een methode voor het bepalen van de grens voldoende/onvoldoende bij studenten. [Above or below par. A method to determine the pass/fail cutoff point in student assessments] [PhD dissertation Rijksuniversiteit Groningen]. Amsterdam: Swets & Zeitlinger, 1971.

47 Muijtjens AMM, Hoogenboom RJI, Verwijnen GM, Van der Vleuten CPM. Relative or absolute standards in assessing medical knowledge using progress tests. *Advances in Health Sciences Education* 1998; **3**: 81-7.

48 Cronbach LJ. How to judge tests: Reliability and other qualities. *Essential of psychological testing*. 5th ed. New York: HarperCollinsPublishers; 1990: 190-222.

49 Brümmer I, Linden ML, Hoogenboom RJI, Scherpbier AJJA, Verwijnen GM. Het oordeel over de Maastrichtse voortgangstoets. Het consumentenoordeel in 1995 [Assessment of the Maastricht progress test. Consumer opinion 1995]. *Bulletin Medisch Onderwijs* 1995; **14**: 174-85.

50 Van der Vleuten CPM. Beyond intuition [Inaugural lecture Maastricht University]. Maastricht: Datawyse, 1996.

51 Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Medical Education* 1983; **17**: 165-71.

52 Frederiksen N. The real test bias. Influences of testing on teaching and learning. *American Psychologist* 1984; **39**: 193-202.

53 Cohen-Schotanus J. Student assessment and examination rules. *Medical Teacher* 1999; **21**: 318-21.

54 Messick S. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 1994; **23**: 13-23.

55 Miller GE. Continuous assessment. *Medical Education* 1976; **10**: 81-6.

56 Van der Vleuten CPM, Scherpbier AJJA, Dolmans DHJM, Schuwirth LWT, Verwijnen GM, Wolfhagen HAP. Clerkship assessment assessed. *Medical Teacher* 2000; **22**: 592-600.

57 Snyder BR. *The hidden curriculum*. 1st ed. Massachusetts: The MIT Press, 1973.

58 Hafferty FW. Beyond curriculum reform: Confronting medicine's hidden curriculum. *Academic Medicine* 1998; **73**: 403-7.

59 Halpin G, Halpin H. Experimental investigation of the effects of study and testing on student learning, retention, and ratings of instruction. *Journal of Educational Psychology* 1982; **74**: 32-8.

60 Ellis JA, Wulfeck II WH, Montague WE. The effect of adjunct and test question similarity on study behavior and learning in a training course. *American Educational Research Journal* 1980; **17**: 449-57.

61 McGuire C. Written methods for assessing clinical competence. In: Hart IR, Harden RM, editors. *Further developments in assessing clinical competence.* Montreal: Heal Publications; 1987: 46-58.

62 McGuire C. Letter to the editor. *Teaching and Learning in Medicine* 1994; **6**: 74.

63 Van Luijk SJ, Van der Vleuten CPM, Schelven RM. The relation between content and psychometric characteristics in performance-based testing. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, editors. *Third International Conference on Teaching and Assessing Clinical Competence.* Groningen, The Netherlands; 1990: 202-7.

64 Van Luijk SJ, Van der Vleuten CPM. A comparison of checklists and rating scales in performance-based testing. In: Hart IR, Harden RM, Des Marchais J, editors. *Current developments in assessing clinical competence.* Montreal: Heal Publications; 1992: 357-62.

65 Verwijnen GM. De student als kwaliteitsbewaker. De rol van de student bij de kwaliteitsbewaking van de Maastrichtse voortgangstoets [The student as quality controller. The student's role in quality assurance of the Maastricht progress test]. *Bulletin Medisch Onderwijs* 1994; **13**: 87-95.

66 Prince CJAH, Visser K. The student as quality controller. In: Scherpbier AJJA, Van der Vleuten CPM, Rethans JJ, Van der Steeg AFW, editors. *Advances in Medical Education.* Dordrecht / Boston / London; 1997: 15-8.

67 Visser K, Prince CJAH, Scherpbier AJJA, Van der Vleuten CPM, Verwijnen GM. Students can be full partners in designing their education. *Academic Medicine* 1997; **72**: 1034-5.

68 Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Schuwirth LWT, Van der Vleuten CPM. Quality assurance in test construction: The approach of a multidisciplinary central test committee. *Education for Health* 1998; **12**: 49-60.

69 Popham WJ, Cruse KL, Rankin SC, Sandifer PD, Williams PL. Measurement -driven instruction: It's on the road. *Phi Delta Kappan* 1985; **66**: 628-34.

70 Blake JM, Norman GR, Keane DR, Mueller CB, Cunnington J, Didyk N. Introducing progress testing in McMaster university's problem-based medical curriculum: Psychometric properties and effect on learning. *Academic Medicine* 1996; **71**: 1002-7.

71 Blake JM, Norman GR, Kinsey E, Smith M. Report card from McMaster: Student evaluation at a problem-based medical school. *Lancet* 1995; **345**: 899-902.

72 Van Berkel HJM, Nuy HJP, Geerlings T. The influence of progress test and block tests on study behaviour. *Instructional Science* 1995; **22**: 317-33.

73 Van Til CT, Van der Vleuten CPM, Van Berkel HJM. Problem-based learning behavior: The impact of differences in problem-based learning style and activity on students' achievement. *Annual Meeting of the American Educational Research association.* Chicago: 1997. ERIC No. TM026783 (ED409333).

74 Van Til CT. Voortgang in voortgangstoetsing. Studies naar de aansluiting van de voortgangstoets op probleemgestuurd onderwijs [Progress in progress testing. Studies on the suitability of progress testing within a problem-based educational context] [PhD dissertation Maastricht University]. Wageningen: Ponsen & Looijen, 1998.

75 Sprooten - Van Hoof MABJ. The measurement of growth in medical knowledge in a non-departmental organized medical school. *Proceedings of the twenty-second annual conference on Research In Medical Education.* USA; 1983: 93-100.
76 Bender W, Cohen-Schotanus J, Imbos T, Versfelt WA, Verwijnen GM. Medische kennis bij studenten uit verschillende medische faculteiten: Van hetzelfde laken een pak? [Medical knowledge in students from various medical schools: Being served with the same sauce?]. *Nederlands Tijdschrift voor Geneeskunde* 1984; **128**: 917-21.
77 Van Hessen PAW, Verwijnen GM. Does problem-based learning provide other knowledge? In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, editors. *Teaching and Assessing Clinical Competence.* Groningen; 1990: 446-51.
78 Albano MG, Cavallo F, Hoogenboom R, Magni F, Majoor G, Manenti F, Schuwirth L, Stiegler I, Van der Vleuten C. An international comparison of knowledge levels of medical students: The Maastricht progress test. *Medical Education* 1996; **30**: 239-45.
79 Van der Vleuten C, Newble D, Case S, Holsgrove G, McCann B, McRae C, Saunders N. Methods of assessment in certification. In: Newble D, Jolly B, Wakeford R, editors. *The certification and recertification of doctors.* Cambridge: Cambridge University Press; 1994: 105-25.
80 Bulte JA, Ket P. Forum 1: Invoeren van de Maastrichtse voortgangstoets in andere faculteiten: J/O/? [Introducing the Maastricht progress test at other medical schools: True/False/I don't know?] *Tijdschrift voor Medisch Onderwijs* 2000; **19**: 31-7.
81 Sambell K, McDowell L, Brown S. "but is it fair?": An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation* 1997; **23**: 349-71.
82 Linden ML, Brümmer I, Hoogenboom RJI, Scherpbier AJJA, Verwijnen GM. De Maastrichtse voortgangstoets. Een vergelijking van drie peilingen van het consumentenoordeel [The Maastricht progress test. A comparison of three samplings of consumer judgement]. *Bulletin Medisch Onderwijs* 1995; **14**: 186-91.

Chapter 3

Quality Assurance in Test Construction The Approach of a Multidisciplinary Central Test Committee

Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Schuwirth LWT, Van der Vleuten CPM
Published in *Education for Health* 1998; **12**: 49-60

Introduction

The objective of testing is to obtain a reliable and valid assessment of students' (cognitive) competencies. In view of the impact of test results on students' progress it is of vital importance to minimize the number of false-positive and false-negative decisions. Not only do test results impact on students' progress, they also influence medical schools' output. In addition, what and how students learn is for a large part directed by tests.[1,2] Thus, careful test construction is of the essence. The construction of good tests requires specific skills and experience, which are not easy to acquire.[3-5] Few teachers view test construction as an exciting task. Fortunately, most teaching institutions do have teachers on their staff who are interested in this aspect of teaching and who are willing to acquire the necessary skills. To ensure fair assessment, it is advisable to separate the roles of teacher and examiner as much as possible. At the Maastricht Faculty of Medicine these considerations have prompted the decision to set up test review committees. These committees are responsible for both quality assurance of the test and the test composition. There are committees for every test with membership comprised of faculty who have shown a special interest in testing. The test committees are chaired by members of the Task Force on Assessment (TFA) of the Faculty of Medicine, which includes physicians and test experts. The TFA's objective is to develop, implement and investigate the assessment system. Over the years other Faculties of Maastricht University have followed this example by setting up similar test evaluation committees. There are also national committees on test assessment, such as the National Board of Medical Examiners in the USA, and the Institut fur Medizinische und Pharmacologische Prüfungsfragen in Germany.[6-8] The 1997 site visit report reiterates the recommendation from the previous report that central test committees be set up. Little has been published on the work of such bodies.[9] To illustrate the workings of test committees, we will describe the approach developed by the Maastricht Progress Test Review Committee (PTRC) and the production process of the progress test.

The progress test

Wijnen introduced the concept of progress testing in 1976 and since the 1977-1978 academic year progress tests have been a fixture in the Maastricht Faculty of Medicine's assessment program.[10] Progress testing is based on the idea that desirable learning behaviour, i.e. self-directed problem-based learning, should not be hampered by assessment. The progress test is therefore aimed at the assessment of any increase in students' medical knowledge (within the framework of the final objectives of undergraduate medical education). The progress test is comprised of questions that a newly graduated MD should be able to answer correctly. A new test, consisting of new and (adapted) previously used items, is constructed for each examination. Four times per academic year all students (year 1 through year 6) take the progress examination simultaneously. The test results reflect how far students have progressed towards the final objective of undergraduate medical education, and thus yield valuable information to guide individual students and evaluate the curriculum.[10,11] For practical reasons (large numbers of both items and students) the test items are of the True / False / Question Mark format.[3] Progress tests contain some 250 items covering fifteen categories derived from the

International Classification of Diseases (ICD).[10] The test blueprint sets out the required number of items for each category.[12] The distribution of the items among the categories is based on morbidity data, the amount of space devoted to the various subjects in general medical textbooks, and a survey among the departments. There are no guidelines on how the items for each category are to be distributed across the different disciplines.

Item bank and item production

The departments devise items that a newly graduated MD must be able to answer correctly. A reference is added to each item by the department to enable students to read background material after the test. All new items are entered into a central, partially computerized item bank, without being evaluated. Items are entered into one of the fifteen categories and they are given a department label and a category label. Some eight months before the date of the progress test, some 400 to 500 items are drawn from the item bank. The items are randomly selected by category in accordance with the distribution set out in the test blueprint. The number of items drawn per category exceeds the required number to ensure that spare items are available in case items are rejected. The number of items that a particular department has contributed to the item bank partly determines how many of that department's items are included in the draw, i.e. the more items, the greater the chance that a particular department is included in the draw. If a department fails to keep its share of items in the item bank at an adequate level, it runs the risk of not being represented in the progress test. To prevent overrepresentation, the items of any one department may take up no more than 6% of the item bank at the time of the draw. If a department exceeds this percentage, surplus items are randomly selected and excluded from the final draw. The items that are drawn are assessed by the PTRC and, after approval, included in the test. Unless items are rejected before or after the test, they remain in the item bank. Items can be re-used after a minimum of three years. As a result the item bank is always changing. Over the years all departments together have contributed a total of approximately 19,000 items. Since 1976 over 6,000 items have been rejected, resulting in some 13,000 items that are currently available for inclusion in tests. After each draw (four times annually), the departments receive an overview listing the number of their items in the item bank by category. Departments are free to add new progress test items to keep their number of items in the item bank at an adequate level.

The Progress Test Review Committee

The PTRC is responsible for constructing the progress tests and for quality assurance. The committee is comprised of the chairperson, deputy chairperson and six members. The chair and deputy chairperson are both physicians and members of the TFA. The members represent the preclinical, clinical and behavioural sciences. The term of membership is four years and can be extended by one additional term. This relatively long term of membership is intended to ensure that full use can be made of the skills and expertise that committee members have acquired on the job. Administrative and logistic support is provided by one staff member, who is also responsible for the item bank and test administration.

Approach used by the PTRC

The PTRC works according to a tight production schedule with times and dates of all its activities being fixed well in advance. All teachers that are involved in item production and quality control are informed of the production dates at the start of every academic year. The total production cycle of one progress test takes thirty weeks (table 1). The PTRC meets twice weekly for three to four hours at a fixed time. The PTRC always has different tests in hand because four progress tests are constructed every year. To coordinate the production schedules of the different tests and to avoid peak loads, good planning is crucial.

Table 1.
Production cycle of one progress test.

week	activity
1, 2	entry into item bank, new items are processed
3	item bank update and item draw
4, 5	assessment by individual members of the PTRC
6-8	first round of 6 PTRC meetings (2/wk)
9	prepare, collect and send correspondence to the departments
10-12	consultations with the departments
13-15	second round of 6 PTRC meetings (2/wk)
16, 17	items with adaptations are processed / preparation item selection by PTRC members
17	check of adaptations and selection of test items (all members of the PTRC)
18	prepare draft test
19	check draft test and report (chairperson and deputy chairperson of the PTRC)
20	prepare final test
21-25	test is printed / logistic preparations for test
26	test (Wednesday 09:00a.m. to 1:00 p.m.)
27	process results and students' comments / preparation of post-test-administration PTRC meeting
28	PTRC meeting / consultations with departments / calculate final scores
29	report to students and committees / preparation of item report
30	item report to departments

Pre-test review

The 400 to 500 items that are drawn from the item bank for each progress test are scrutinized by the PTRC. Each committee member is responsible for the items pertaining to one or more (the maximum is four) categories of the test blueprint. Committee members first review the items of their assigned categories individually. The results of these initial individual reviews are then discussed and checked in plenary sessions of the committee to provide feedback to one another. A first round of six meetings is reserved for these sessions. The committee meets in a room where a large number of books can be consulted to check items' content validity. The review takes account of whether the item is formulated clearly and unequivocally, whether it is correct contentwise, and whether it is

adequately and unequivocally documented. If an item fails to meet any of these requirements concrete editorial changes are suggested. In these cases the contact person of the department that supplied the item is consulted either in writing or face to face. An item can only be adapted or removed after consultation with the department concerned, because it is the departments that are responsible for the quality of item content. A second round of six plenary committee meetings is planned after the consultations with the departments. These sessions may result in approval or removal of an item, or in more consultations. Since many items require more than one round of consultations and since the committee is working on different progress tests simultaneously, the PTRC is usually working on a thousand items at any one time.
A final meeting of the PTRC is planned to compose the draft PT. Every committee member makes a final draft for their own categories. During this stage, members check whether all adaptations have been processed accurately and a selection is made from the available approved items. How many items a department eventually contributes to the test depends both on the department's initial number of items in the item bank and its cooperation with the PTRC in test production. A department that fails to respond adequately to the PTRC's comments runs the risk that fewer of its test items are available for inclusion in the progress test. Every member of the PTRC selects a small surplus of items for the final draft, in case items need to be removed to achieve an even distribution of keys and sciences clusters (i.e. basic, clinical, and behavioural). The final draft of the complete progress test is again checked and corrected by the chairperson and the deputy chairperson. Originally, this was intended as a final check for typing errors and overlap between items, but experience has since taught that even at this stage serious problems are detected. Thus, nowadays this final round includes a check for incorrect wording, references, and content-related errors. The results of this check are reported in writing to the PTRC members and discussed in a plenary committee meeting by way of intervision. Subsequently, key distribution (proportion correct/incorrect) and cluster distribution are determined. By removing items the committee aims at an even distribution of keys (50% correct, 50% incorrect) and the desired distribution among clusters (40% basic subjects / 40% clinical subjects / 20% behavioural sciences subjects). The chairperson performs a final check of the accuracy of all adaptations and the test is then ready to be printed.

The majority (over 80%) of the items drawn from the item bank (both new items that have never been assessed and items used in previous tests) are not suitable for inclusion in the progress test without adaptation. In almost all cases (97%) adaptation concerns the way the item is formulated. Over half (56%) of the adaptations also involve content-related changes.

Posttest review
After the test the items are reviewed again. Despite the careful pretest review procedure, some of the included test items prove to be inadequate. Both the departments and the PTRC overlook some problematic items. Students can alert the PTRC to these items. To this end, students are given the opportunity to participate in the assessment process.[13,14] They can offer comments until one week after the test. Only typewritten comments are acceptable to avoid readability problems. When the student includes references in his or her comments,

copies of the referenced texts must be enclosed. On average 4% of students comment on some 20% of the items in one single test.

To trace problematic items, also statistical parameters of the item results are used, the item analysis.[3] The students' comments and the results of the item analysis are added to a text file containing the item texts. The chairperson and deputy chairperson judge every item taking account of the item analysis and students' comments. Students' comments are evaluated by checking the references provided. If students' arguments are valid, the item is put on the agenda of a plenary PTRC meeting. The agenda for this meeting also includes all items with an unusual answering pattern, i.e. no increase in the number of correct answers over years 1 through 6, a striking peak or drop in a particular class, an item that is answered incorrectly by over 30% of sixth-year students or that is not answered by more than 50% of sixth-year students. Ten days after the test, these items are discussed in this final plenary PTRC meeting. Committee members receive the items together with each item's history and the chairperson's and vice chairperson's recommendations as to how it should be dealt with. Discussions should result in decisions on withdrawing an item or changing the key. If the PTRC considers eliminating an item or changing its key, the department that contributed the item is consulted. The department and the committee must reach agreement on whether to leave the item as it is, change the key, or withdraw the item. No item is removed or key changed without permission from the department that has contributed the item. After this consultation round with the departments, the definitive test scores are calculated. The items that have been withdrawn are not considered in calculating the scores. An average of 5.5% of items are withdrawn, but this percentage varies substantially across tests with percentages ranging from 0.4% to 11.5%. The number of key changes averages 0.7% ranging from 0% to 2%. After the key changes have been incorporated and the eliminated items removed, there always remain items that relatively few students have answered or that relatively many students have answered incorrectly. At the end of the sixth year (test 24) an average of 15% (some 34 items) of items is answered by less than 50% of sixth-year students. Progress tests show considerable differences in this respect, however. The number of items that are answered incorrectly by more than 30% of sixth-year students remains fairly stable at around 16% (some 36 items).

Reporting

The results are reported to the Certifying Committee, which formally establishes the results and announces them to the students. The students receive a form listing both their own individual score and the mean score of their class. The scores are given by category and by discipline and are expressed as percentages of the highest possible scores. The form lists both the correct/incorrect/? scores and the correct-minus-incorrect score. In addition to the individual score a "+" or "-" indicates whether the student has scored high or low compared with the entire class. At the bottom of this list the overall score (correct-minus-incorrect) and the qualification (pass/fail) are entered. The form also states which items were withdrawn and which keys were changed (figure 1). The upper half shows the individual and the class score by category, the lower half shows the scores by department.

Figure 1. Results form sent to students.

RESULTS BY CATEGORY (Percentages)													
Category	no.of items	individual				class (n=215)							
		Correct	Incor.	?	C-I	Correct	SD	Incor.	SD	?	SD	C-I	SD
1 Respiratory	28	50++	7--	43	43++	37	12	23	10	41	16	14	16
2 Hematol and lymph	14	36	14	50	21	43	15	14	12	43	20	28	19
3 Muscle and skeleton	18	44+	22++	33--	22	35	13	11	10	54	17	23	15
4 Mental health	17	53-	29++	18	24--	62	14	13	9	25	14	49	19
5 Reproduction	16	13-	31+	56+	-19--	41	15	20	12	38	18	21	20
6 Cardiovascular	28	71++	14	14--	57++	50	13	16	9	34	16	34	15
7 Hormone and metabol	16	19-	13	69+	6	28	17	16	11	56	21	11	19
8 Skin and conn. tissue	10	50+	0-	50	50++	31	19	13	14	56	24	17	22
9 Social factors illness/health	14	64	14	21	50	58	14	16	11	26	16	43	18
10 Gastrointestinal	23	13-	4-	83++	9	24	13	14	11	62	20	9	13
11 Renal and urinary	18	11-	17	72+	-6-	23	15	16	10	62	19	7	16
12 Nervous and sensory	21	33	14-	52	19	35	13	21	11	44	18	15	16
13 Other	11	9--	18	73++	-9--	39	17	12	12	49	19	27	22
14 Research and methodology	9	56+	22	22-	33	44	16	15	14	41	22	29	21
15 Indiv. factors illness/health	6	50	50++	0--	0--	57	20	22	15	21	20	35	29
Total	249	39	16	45	22	39	9	16	7	44	14	23	8

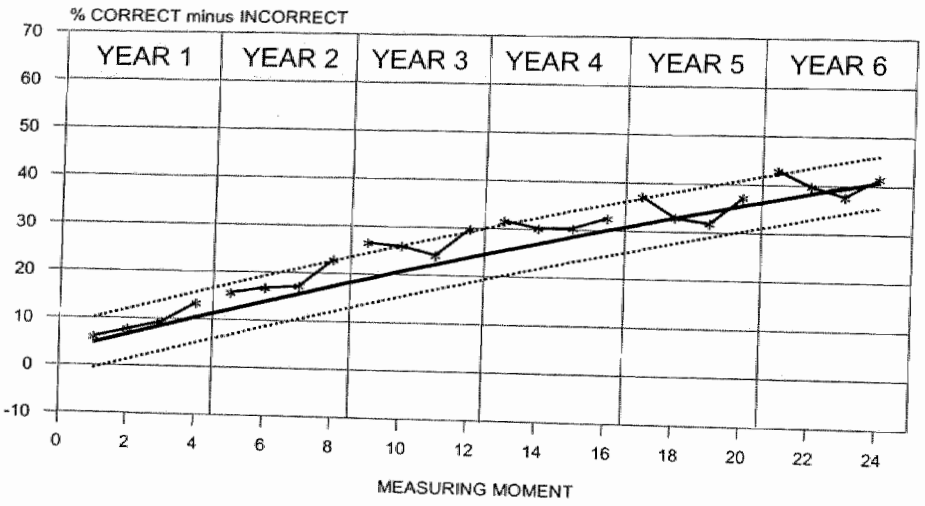
RESULTS BY DEPARTMENT (Percentages)													
Department	no. of items	individual				class (n=215)							
		Correct	Incorrect	?	C-I	Correct	SD	Incor.	SD	?	SD	C-I	SD
1 Anatomy/Embryology	10	30-	20	50+	10-	45	16	19	13	37	19	26	22
2 Biochemistry	15	33	7-	60	27+	31	18	16	11	53	22	15	20
3 Biophysics	0	-	-	-	-	-	-	-	-	-	-	-	-
4 Pharmacology	10	50+	20	30-	30	37	18	18	13	45	20	18	24
5 Physiology	20	55	5--	40+	50++	48	14	23	12	30	16	25	20
6 Genetics/Cell biology	10	0--	30+	70++	-30--	49	18	20	13	31	21	29	23
7 Immunology	9	22-	11+	67+	11--	42	20	4	8	54	23	39	21
8 Medical microbiology	18	6--	11	83++	-6-	25	13	16	12	59	20	9	15
9 Pathology	9	33	11	56	22	30	17	13	12	57	20	17	21
Total basic sciences	101	30-	13-	57+	17	38	11	17	7	45	15	21	10
10 General surgery	10	60++	10	30-	50++	34	17	14	13	52	23	19	20
11 Cardiology	12	67++	8-	25	58++	49	16	16	11	34	19	33	20
12 Dermatology	9	44+	0-	56	44++	30	20	12	15	58	25	19	24
13 Obstetrics/Gynecology	10	20-	50++	30-	-30--	34	17	24	15	42	21	9	24
14 Family medicine	10	40	20	40	20-	46	17	15	11	40	19	31	21
15 Internal medicine	14	14-	7-	79+	7	24	15	18	13	58	21	6	18
16 Pediatrics	9	33	22	44-	11	26	17	17	14	57	24	8	19
17 Ear , Nose , Throat	5	80++	0--	20-	80++	39	22	18	16	43	25	21	30
18 Neurology	5	20	20	60	0	30	23	23	20	47	27	8	33
19 Ophthalmology	4	25	25+	50	0	22	22	14	17	63	26	8	29
20 Orthopedics	3	0	33++	67-	-33--	8	21	4	11	88	29	4	17
21 Pulmonology	9	33	11	56	22	38	16	13	12	49	20	24	21
22 Radiology	3	67++	33	0-	33+	32	27	43	27	25	26	-11	48
23 Rehabilitation	1	0	0	100+	0	15	36	12	32	73	44	3	51
24 Urology	2	0-	0	100+	0-	29	33	3	12	68	35	26	36
Total clinical sciences	106	38	16	46	22+	33	10	17	8	50	16	17	9
25 Health care economics	2	50	50	0-	0	35	30	42	37	23	30	-7	60
26 Epidemiology	4	50	0-	50+	50+	48	23	18	20	34	25	30	36
27 Public health law	7	71	14	14	57	73	15	9	11	18	15	64	22
28 Medical ethics	5	80++	20	0--	60+	52	22	13	15	35	24	40	30
29 Medical psychology	10	70++	30++	0--	40	51	14	16	13	33	18	35	20
30 Medical sociology	5	40--	60++	0-	-20--	70	20	15	17	14	17	55	33
31 Psychiatry	9	56	11	33	44	61	16	11	11	29	17	50	21
Behavioral sciences	42	62	24++	14-	38	58	10	15	7	27	13	43	12
Total	249	39	16	45	22	39	9	16	7	44	14	23	8

- / -- / + / ++ / low, high compared with group

The cutoff score for this test is: 15.85%
Your CORRECT - INCORRECT score is: 22.49% (Absolute: 56) Your result: PASS
Eliminated items: 38,68,69,133,152,155, 242. Changes of key: 12,116,215, 243.

The results are also reported to the Education Committee, the TFA, and all departments. All bodies involved receive a list of the mean scores by class for the entire test and by categories, departments and departmental clusters that contributed the items. The report includes the so-called growth curves and the results on the progress tests of the current academic year (figure 2).

Figure 2.
Growth curve of results on the 4 annual progress tests in the period September 1985 - May 1995. The curve represents the "line of best fit" for the mean scores of 10 classes (10 times 24 data points). The dotted lines represent the 95% confidence intervals. The asterisks indicate the mean scores of the six classes on 4 consecutive tests.



In addition, every department receives an item report for all of their items included in the test. For each item the definitive text, the answering profile, supplementary item-analysis parameters, any student comments, and history (previous versions, results of previous tests, any changes in the key, withdrawn after the test, etc.) are reported. Feedback is complemented by PTRC comments. The latter comprise a brief interpretation of the item analysis and possible explanations for problems identified. The item report is available to students (figure 3).

Figure 3.
Example of an item report.

PROGRESS TEST FACULTY OF MEDICINE –UNIVERSITY OF MAASTRICHT: December 1994 ***TASK FORCE ON ASSESSMENT***

PHYS 0134 /06-00195
Systolic coronary circulation rate differs from diastolic coronary circulation rate.
105- The diastolic circulation rate is higher.
KEY: True

ITEM ANALYSIS:

YEAR:	1	2	3	4	5	6
%-CORRECT:	11	58	58	64	71	79
%-INCORRECT:	8	21	28	20	20	15
%-?:	82	21	14	16	9	6
RIT Cml:	0.089	0.001	0.118	0.104	0.093	0.216
DI Cml:	-0.037	-0.026	0.048	0.133	0.026	0.160

STUDENT COMMENTS:
#93152
Coronary circulation rate varies considerably during the heart cycle.
The left coronary artery is compressed by the high intraventricular pressure during systole (esp. the inner capillaries and veins). Circulation rate here is highest during diastole. The right coronary artery is less affected by the intraventricular pressure, because it is located lower compared with the left coronary artery, i.e. coronary circulation rate follows aortic pressure and is therefore highest during systole. In 50% of people the right coronary artery is dominant (in 20% the left artery is dominant, in 30% there is no dominance) - i.e. the right coronary artery serves both the right-hand side and part of the left-hand side. This implies that in 50% of people the coronary system consists for the greater part of the right coronary artery, where systolic circulation rate is higher than diastolic circulation rate.
Lit.: Bernards en Bouman, Fysiologie van de mens, 1988, blz. 361.
#XXXX8
I think this is a questionable item because coronary circulation cannot be viewed as one single entity. Early during diastole the circulation rate is indeed higher in the left coronary artery whereas the systolic circulation rate is very low. HOWEVER, the right coronary artery follows aortic pressure, i.e. the maximum circulation rate occurs at the end of the systole.
(Bernards & Bouman 1988, pag. 361)

COMMENTS OF THE PROGRESS TEST REVIEW COMMITTEE (PTRC):
After consultation with the contact person from the department it was decided to WITHDRAW the item.
Motivation: The students' comments are correct.
Do you think the item should be revised or should it be removed from the progress test item bank?

History:
Original version (used in PT-DEC87: response comparable):
Diastolic circulation rate is higher than systolic circulation rate [true]

Appeal procedure
After the test results have been published, a student who disagrees with the way the result was arrived at, can appeal to the College of Appeal for the Examinations. The right of appeal is statutory. If a student appeals, the chairperson of the PTRC provides the

Certifying Committee with written comments and advice with respect to the appeal, if necessary, after consultation with the experts involved. It is a statutory requirement that these comments and advice are discussed with the student to examine whether an amicable settlement can be reached. If no agreement is reached, the appeal is heard at a public session of the College of Appeal for the Examinations, and expert witnesses may be called. Since 1990, appeal cases have occurred fairly regularly. Before that only one appeal case occurred, presumably because until March 1990 the possibility of appeal was not mentioned in the official test result form. Until then, the option was only included in the study guide. In seven years (1990-1996), 121 items from 19 out of the total of 28 tests have led to appeal cases. This is a relatively small number of items, i.e. it is less than 2% of the total number of items included in tests during these 7 years. The majority of appeals were resolved by amicable settlement. On 27 of 121 items the student and the Examination Committee failed to reach agreement and the Board of Appeal for the Examinations had to decide. In most cases students appeal because they stand to gain if their test result is changed. Over the past 7 years, 59 different students have appealed, most of whom needed a slightly improved score to achieve a better qualification.

Discussion

It is the task of test review committees to assess independently of the item authors whether test items are valid indicators of knowledge. By identifying potential error sources the committee strives to minimize the risk of false-positive or false-negative decisions and ensure that the test does indeed measure what needs to be measured. It appears to be worthwhile to have a multidisciplinary group scrutinize test items to identify problems relating to both form and content of the items. The approach adopted by the Maastricht PTRC to enhance test quality might be regarded as a form of peer review. It is an approach, which has evolved over the past twenty years and which is continually being adapted to the prevailing circumstances, demands, and ideas. A recent development is the publication of a national document, which sets out the final objectives of undergraduate medical education in the Netherlands. Currently, proposals are being considered to tailor test items to this document.[15] There are also plans to improve efficiency by (re)computerizing the procedure.

From our experiences we have learned that the described approach to test review leads to timely detection and solution of problems. The majority (over 80%) of items drawn from the item bank are not immediately suited for inclusion in the final progress test. Also items used in previous tests often prove to need adaptation before they can be re-used. Despite the intensive review procedure, student comments and item analysis identify an average of 5.5% of items that should be withdrawn from the test. This percentage has remained stable over the past twenty years. The same applies for the psychometric parameters. Over the years test reliability, mean class scores, and percentage of items that are not answered or answered incorrectly have remained stable.[10,16]

There is no proof that the approach described in this article produces better test items. What it does provide, however, is a clear, thorough and open assessment procedure, which signifies to all those involved that the test is taken seriously. Given that test results have far-reaching consequences for students' progress, students are entitled to careful quality assurance of tests. Peer review is commonly used in evaluating research, but its application in education is rare. Based on our experiences in test review, we believe that peer review could also be valuable in educational quality assurance.

References

1 Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Medical Education* 1983; **17**: 165-71.

2 Frederiksen N. The real test bias. Influences of testing on teaching and learning. *American Psychologist* 1984; **39**: 193-202.

3 Ebel RL, Frisbie DA. *Essentials of educational measurement*. 5th ed. New Jersey: Englewood Cliffs, 1991.

4 Haladyna TM, Downing SM. A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education* 1989; **2**: 37-50.

5 Cox KR. How to construct a fair multiple choice question paper. In: Cox KR, Ewan CE, editors. *The medical teacher*. 2nd ed. London: Churchill Livingston; 1988: 157-60.

6 Stokes JF. Examining in the united states: The national board of medical examiners. *British Journal of Medical Education* 1967; **1**: 320-9.

7 Hubbard JP. *Measuring medical education*. 2nd ed. Philadelphia: Lea & Febiger, 1978.

8 Kraemer HJ, Duppré HJ, Boelcke G, Micaelis J, Voigtmann K. *Institut für medizinische und pharmazeutische prüfungsfragen. Aufgaben, entwicklung, analysen*. Mainz: Verlag Druckhaus Schmidt & Bödige, 1976.

9 Downing SM, Haladyna TM. Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education* 1997; **10**: 61-82.

10 Van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher* 1996; **18**: 103-9.

11 Van der Vleuten CPM, Scherpbier AJJA, Wijnen WHFW, Snellen HAM. Flexibility in learning: A case report on problem-based learning. *International Higher Education* 1996; **2**: 17-24.

12 Verwijnen M, Imbos T, Snellen H, Stalenhoef B, Pollemans M, Van Luyk S, Sprooten M, Van Leeuwen Y, Van der Vleuten C. The evaluation system at the medical school of maastricht. *Assessment and Evaluation in Higher Education* 1982; **7**: 225-44.

13 Bandaranayake RC, Cox KR. Writing multiple choice questions. In: Cox KR, Ewan CE, editors. *The medical teacher*. 2nd ed. London: Churchill Livingston; 1988: 152-6.

14 Prince CJAH, Visser K. The student as quality controller. In: Scherpbier AJJA, Van der Vleuten CPM, Rethans JJ, Van der Steeg AFW, editors. *Advances in Medical Education*. Dordrecht / Boston / London; 1997: 15-8.

15 Metz JCM, Stoelinga GBA, Pels Rijcken - van Erp Taalman Kip EH, Van den Brand - Valkenburg BWM. *Blueprint 1994: Training of doctors in the netherlands. Objectives of undergraduate medical education*. Nijmegen: University Publication Office, 1994.

16 Swanson DB. *Review of the assessment system used by the university of limburg medical school*. [EPG-publ.nr. 88-22] Maastricht: University of Limburg; 1988.

Chapter 4

The progress test
Assessing the objectives of undergraduate medical education

Verhoeven BH, Buijs C, Scherpbier AJJA, Verwijnen GM
Under editorial review

Introduction

During the past decades the objectives of undergraduate medical education have been described in many countries [1,2]. In the Netherlands they were published in 1994 in a document entitled Blueprint 1994: training of doctors in the Netherlands [3]. Although it is a good thing for medical schools to have well defined and clearly described objectives, we also know that the curriculum on paper is not always the same as the learned curriculum [4,5]. Or, in other words, inclusion of a subject in the curriculum does not necessarily guarantee that students attain adequate competence. For clinical clerkships it has been shown that students' exposure to clinical experiences varies [6]. As a consequence the Blueprint's objectives are not always being met [7].

In the Maastricht medical school students' achievement of the end objectives is evaluated by the progress test. Since 1976, the progress test has been administered four times a year to all students regardless of their class. Each progress test is a comprehensive examination constructed with the intention to reflect the final objectives of the curriculum [8]. To ensure that tests are equivalent, a blueprint is used, derived from the International Classification of Diseases. In the blueprint, each discipline is assigned to one of three clusters, namely basic sciences (anatomy, biochemistry, pharmacology, physiology, genetics & cell biology, immunology, microbiology, and pathology), clinical sciences (surgery, cardiology, dermatology, obstetrics and gynaecology, family medicine, internal medicine, paediatrics, ENT, neurology, orthopaedics, ophthalmology, pulmonology, radiology, rehabilitation medicine, and urology) and behavioural/social sciences (health care economics, epidemiology, health care law, ethics and philosophy, medical psychology, medical sociology, and psychiatry). The department/discipline that produces the item determines the cluster where the item belongs to. An item that is written by the surgical department concerning anatomy is classified in the cluster of clinical sciences. The progress test contains approximately 250 items in the (multiple) true / false / I do not know format. Test items are supplied by the departments and item producers are asked to write items that test knowledge at the level of a newly graduated medical doctor. Thus it may be assumed that the progress test can be regarded as an operationalisation of the medical school's final objectives. In an earlier article the production process of the progress test and the quality assurance procedure were described [9].

The study reported in the present article investigated the match between the objectives stated in Blueprint 1994 and progress test items. We present a method for comparing item content and level with Blueprint objectives as well as the results of a pilot in which the method was tested on a randomly selected progress test. Finally, we present students' results on the progress test studied and evaluate them in light of the correspondence between item content and Blueprint objectives.

Method

Firstly, we wanted to devise a method whereby we could determine whether the content of a test item corresponded to any of the Blueprint objectives. In Blueprint 1994 the objectives of undergraduate medical education have been operationalised as general objectives, problems as starting points for training, and discipline-related objectives. We first looked at the general objectives. For a randomly selected series of items from a randomly selected progress test the authors tried to identify links between test items and general objectives. Soon it became obvious that one way or another all test items reflected the general objectives. This was due to the very broad way in which the general objectives have been defined. For instance, virtually any medical subject is in some way related to the definition "man in somatic, mental and social respect". Another drawback of the general objectives is the absence of defined levels of competence. Secondly, we looked at the 'problems as starting points for training'. Since no corresponding competence levels are provided for this operationalisation of the Blueprint objectives either, the 'problems' were not suitable for our study. This left only the discipline-related objectives against which test items could be matched. Fortunately, the discipline-related objectives are not only described clearly and in detail but they are also accompanied by a precise description of the required competence level. For example, 'acute glaucoma' requires competence at level D (the doctor must make the diagnosis personally). The conclusion was that the discipline-related objectives were a suitable benchmark to determine whether test items reflected Blueprint objectives.

Before describing the four-step method we developed for this study, we will briefly describe the structure of progress test items and the operationalisation and the levels of competence of the discipline-related objectives in the Blueprint. Progress tests are knowledge tests and consist of statements which may be linked to a stem which provides necessary information. Students are asked to indicate whether the statements are true or false. They may also use the do not know option, which is penalized nor rewarded. The test result is expressed as the percentage correct minus incorrect answers.

In the Blueprint 1994 discipline-related objectives consist of lists of clinical pictures and skills for the different disciplines. For every clinical picture the competence level is indicated. Four competence levels have been defined. In the definitions, "doctor" refers to recent medical graduates. 1. be able to recognise or place, i.e. the doctor must have heard of it, but does not have to be able to deal with it. 2. be able to cope with in clinical practice, i.e. in a real situation the doctor must be able to consider it as a diagnosis. D. The doctor must make the diagnosis personally. T. The doctor must carry out the therapy personally.

The four-step approach
We developed a four-step method to capture in a quantitative manner the association between test items and objectives. The method is described below and illustrated by an example of how the method is applied to classify a test item.
Step I. determine the subject of the progress test item.

Progress test item: In cases of hyperventilation the arterial pCO_2 changes. The arterial pCO_2 increases (true or false).
The subject of the test item is hyperventilation.

Step II. determine the level of competence needed to give the correct answer.
Regarding the test item on hyperventilation the following was considered.
The question concerns the practical management of the concept of hyperventilation. The item is not related to diagnostics or therapy. Thus the item requires competence at level 2.

Step III. trace the subject of the item in Blueprint 1994.
Four types of connection were distinguished: items were associated with the Blueprint by a direct connection, an indirect connection, a very indirect connection or they were unrelated. When the subject identified in step I or another subject derived from the test item corresponded to one of the discipline-related objectives, there was a direct connection. When no corresponding objective was found, but the item was related to a clinical picture, there was an indirect connection. The connection was very indirect, when the item could only be matched with one of the other clinical pictures listed under the discipline(s) that linked with the item, for this the whole clinical pictures list could be searched.
The connection between the test item on hyperventilation and the Blueprint was a direct one, because hyperventilation is included in the discipline-related objectives.

Step IV. congruence between the level of competence required for the test item and that required in the Blueprint.
Using the competence level indicated in the Blueprint as the benchmark, four degrees of congruence were distinguished: good, moderate, poor and no congruence. Congruence was said to be good when the level required to give the correct answer for the item (step II) was the same as or lower than the level indicated in the Blueprint. When the test item required a higher knowledge level than that indicated in the Blueprint, congruence was moderate (too difficult by one level) or poor (a difference of two or more levels).
The Blueprint requires level T for hyperventilation. This should be amply sufficient to answer the question correctly, i.e. congruence is good.

Figure 1 illustrates the four steps with a different test item. After developing the method and having gained some experience in its application, the authors analysed a complete progress test. The authors went through the steps individually and discussed the resulting classifications afterwards. Disagreement was resolved by discussion until consensus was reached.

Figure 1.
Example to illustrate the four-step method to identify whether a test item corresponds to the objectives of Blueprint 1994.

Secretion of HCI through the parietal cell into the stomach can be suppressed in different ways. A) through a H_2 receptor block. B) Through H^+/K=ATP-ase inhibition. Some anti-ulcerative drugs act as indicated under A. The following drug(s) fall(s) into that category:
- ranitidine
- cimetidine
- omeprazole

Step I:	determine the subject of the test item: *anti-ulcerative drugs*
Step II:	determine the level of competence required to answer the test item: *T (= The doctor must carry out the therapy personally)*
Step III:	trace the subject in the Blueprint: *"anti-ulcerative drugs" does not occur in the discipline related objectives; no other subject from the test item could be traced either; a related clinical picture was found, however: ulcus ventriculi/duodeni. This means that there is an indirect connection.*
Step IV:	congruence between the competence level required for the test item and the level indicated in the Blueprint: *ulcus ventriculi/duodeni requires competence at level 2 in the Blueprint (= be able to cope with in clinical practice). The level required to answer the test item correctly is T. This means that the item is too difficult by two levels, i.e. congruence is poor.*

Results

Connection between progress test items and Blueprint 1994
The progress test that was analysed for the study (May 1994) consisted of 230 test items. Despite having become quite adept at classifying items, it took the authors several sessions lasting more than six hours in all to analyse the items and reach consensus. There were frequent disagreements, although generally over minor details and consensus could generally be reached very quickly. In table 1 the connection between progress test items and the Blueprint is presented. 128 (55.7%) test items were classified as having a direct connection with Blueprint 1994. There was an indirect connection for 43 (18.7%) items, a very indirect connection for 8 (3.5%) items and no connection for 51 (22.2%) items. In all, a connection was found for 179 test items.

Table 1.
Number of progress test items by type of connection and the distribution of the items over the three sciences clusters in the progress test.

	Cluster											
	Basic sciences			Clinical sciences			Behavioural sciences			Total		
Connection	n	%	(%)	n	%	(%)	n	%	(%)	n	%	(%)
Direct	36	28	(39)	74	57	(74)	18	14	(47)	128	100	(56)
Indirect	25	58	(27)	18	41	(18)	0			43	100	(19)
Very indirect	6	75	(7)	2	25	(2)	0			8	100	(4)
Connected	67	37	(73)	94	52	(94)	18	10	(47)	179	100	(78)
Unrelated	25	49	(27)	6	11	(6)	20	39	(53)	51	100	(22)
Total	92	40	(100)	100	43	(100)	38	16	(100)	230	100	(100)

As described in the introduction, progress test items are distributed over three clusters: basic sciences, clinical sciences and behavioural/social sciences (table 1). The majority of the 179 connected test items fell into the category of clinical sciences (52.5%). Of the 51 unrelated test items, 25 (49%) concerned basic sciences, 6 (11.8%) clinical sciences and 20 (39.2%) were related to behavioural sciences.

On closer inspection some of the unrelated clinical sciences items were found to belong to a different category. Four of those items were about anatomy, and thus were really basic sciences items. One of the six items asked about costs, which puts it in the category of behavioural sciences. The majority of the unrelated test items in the category basic sciences required knowledge of specific (anatomical) structures or basic sciences research techniques. The unrelated items did not describe patient problems, whereas most of the connected test items did do so. The twenty unrelated test items about behavioural sciences covered a variety of subjects including theory of sciences, epidemiology, statistics, legislation, economics and sociology. Half of the items were presented in a (medical) practice-related context. Nevertheless, the subjects did not figure among the discipline-related objectives. In order to be able to give a correct answer to the other unrelated items, detailed knowledge was required of general (largely other than medical) subjects, which were not found in the Blueprint. The eighteen connected items about behavioural sciences dealt with psychiatric or medical-psychological subjects. Clinical sciences items showed the highest percentage of connections with Blueprint 1994. Of 100 clinical sciences items, 94 were connected with the Blueprint. Of 92 basic sciences items 67 (72.8%) were connected and of 38 behavioural sciences items only 18 (47.4%) were about subjects that could be traced in Blueprint 1994.

Congruence between the competence levels required for the test items and the Blueprint objectives

For 123 (53.5%) test items the level required to give a correct answer corresponded to the level indicated in Blueprint 1994. For 34 (14.8%) test items congruence was moderate and for 22 (9.6%) it was poor. The remaining 51 (22.2%) test items were unrelated to Blueprint

1994 and therefore no congruence was determined. The highest congruence was found for test items about clinical sciences (table 2).

Table 2.
Number of progress test items by congruence of competence level and the distribution of the items over the three sciences clusters in the progress test.

	Cluster											
	Basic sciences			Clinical sciences			Behavioural sciences			Total		
Congruence	n	%	(%)	n	%	(%)	n	%	(%)	n	%	(%)
Good	45	37	(49)	65	53	(65)	13	11	(34)	123	100	(53)
Moderate	9	27	(10)	22	65	(22)	3	9	(8)	34	100	(15)
Poor	13	59	(14)	7	32	(7)	2	9	(5)	22	100	(10)
None	25	49	(27)	6	12	(6)	20	38	(53)	51	100	(22)
Total	92	40	(100)	100	44	(100)	38	17	(100)	230	100	(100)

Connection and congruence combined
Of 230 test items 85 (37.0%) demonstrated a direct connection and good congruence. For 27 test items (11.7%) the connection was direct and congruence moderate and 16 items (7.0%) had a direct connection and poor congruence. Of the test items which were indirectly related to the Blueprint 30 (13%) had good congruence, 7 (3.0%) showed moderate congruence and for 6 (2.6%) congruence was poor. The eight very indirectly connected test items demonstrated good congruence (table 3).

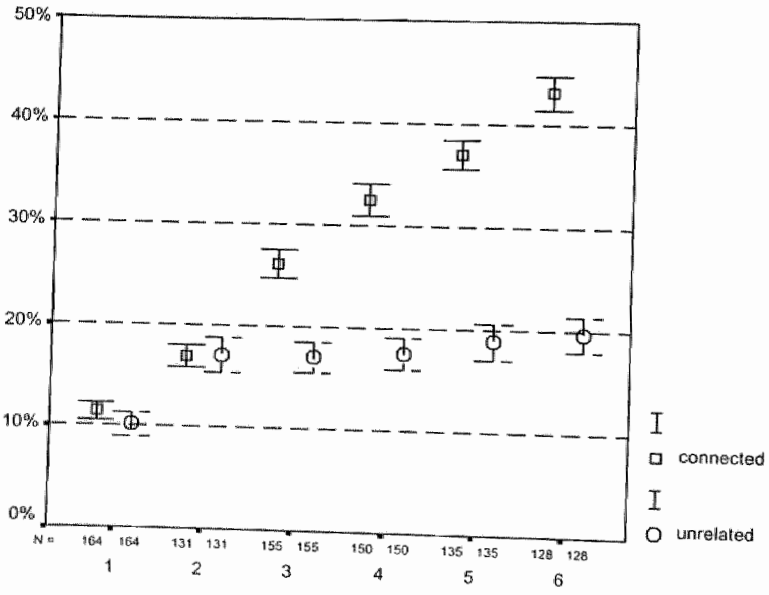
Table 3.
Number of progress test items by combinations of connection and congruence.

	Connection									
	Direct		Indirect		Very indirect		Unrelated		Total	
Congruence	n	%	n	%	n	%	n	%	n	%
Good	85	37	30	13	8	4	0		123	
Moderate	27	12	7	3	0		0		34	
Poor	16	7	6	3	0		0		22	
None	0		0		0		51	22	51	
Total	128		43		8		51		230	100

Progress test results and the connection between item and Blueprint

After analysing the progress test using the four-step method, we wanted to explore the relationships between how well students did on an item and its connection and congruence with Blueprint 1994. Figure 2 shows the mean outcomes and 95% confidence intervals of six cohorts of students for the connected and unrelated items. The most striking finding were the low scores on unrelated test items. Starting in year three, scores on unrelated items are markedly lower than those on connected items. It is also interesting that the scores on the unrelated items did not increase after year 2. Students in third year or higher obtained scores on items connected with Blueprint 1994 that were slightly above the mean scores from previous progress tests. The increase in scores with increasing years of study was similar to that found for progress tests in general [8,10].

Figure 2.
The mean percentages scores (correct minus incorrect) and 95% confidence intervals for 179 items that were connected with the Blueprint and for 51 unrelated items. The horizontal axis represents the six student cohorts. The total number of students who took the test was 863. Below the X-axis the number of students in the 6 cohorts is reported.



In table 4 the mean correct, incorrect, correct minus incorrect as well as "I do not know" scores of the sixth year students are presented on the test in total and the two sub-tests (connected versus unrelated). The worse test results for the Blueprint-unrelated subtest is mainly caused by a higher percentage of "I do not know" answers and slightly more

incorrect answers. Sixth year students score higher on the Blueprint-connected sub-test compared to the average (normal) progress test score at the end of undergraduate medical education [10]. The score is even slightly higher than that of graduated doctors on a normal progress test [8,11].

Table 4.
The mean correct, incorrect, correct minus incorrect and "I do not know" scores of the sixth year students on the test in total, the two sub-tests and on average over 20 years of progress testing[10].

	mean test score of sixth year students (%)			
	connected	unrelated	total	average
correct	61	41	56	58
incorrect	17	21	18	17
I don't know	22	37	25	25
correct minus incorrect	43	20	38	41

In routine post-test review, items with an unusual answer pattern are selected, according to certain criteria [9], to find items that have a potential content validity problem. Items that are not answered by more than 50% of sixth year students, and/or answered incorrectly by more than 30% of sixth year students are among them. In table 5 the percentage of items of the connected and unrelated sub-tests which satisfy these criteria are given. Of the unrelated sub-test three times as many questions are not answered by sixth year students compared to the connected sub-test. At the same time a double number of items is answered incorrectly.

Table 5.
The percentage of items that are not answered by more than 50% of sixth year students and answered incorrectly by more than 30% of sixth year students for the test in total, the two sub-tests and on average over 20 years of progress testing[10].

	% of items			
	connected	unrelated	total	average
> 50% of Year 6 do not know	10	31	14	15
> 30% of Year 6 incorrect	13	26	17	16

Progress test results and the level of connection and congruence between item and Blueprint

The results of the connected items are further analysed by splitting the test into levels of connection and levels of congruence. Figure 3a shows the mean outcomes and 95% confidence intervals of six cohorts of students on the three levels of connection (direct,

indirect, very indirect) and figure 3b depicts the three levels of congruence (good, moderate, poor). The former shows that students score worse on very indirect connected items in Year 4 to 6. This concerns eight questions. The latter reveals no unequivocal relationship between the level of congruence and test results.

Figure 3a.
The mean percentages scores (correct minus incorrect) and 95% confidence intervals for the three levels of connection with the Blueprint: direct (128 items), indirect (43 items), and very indirect(8 items).

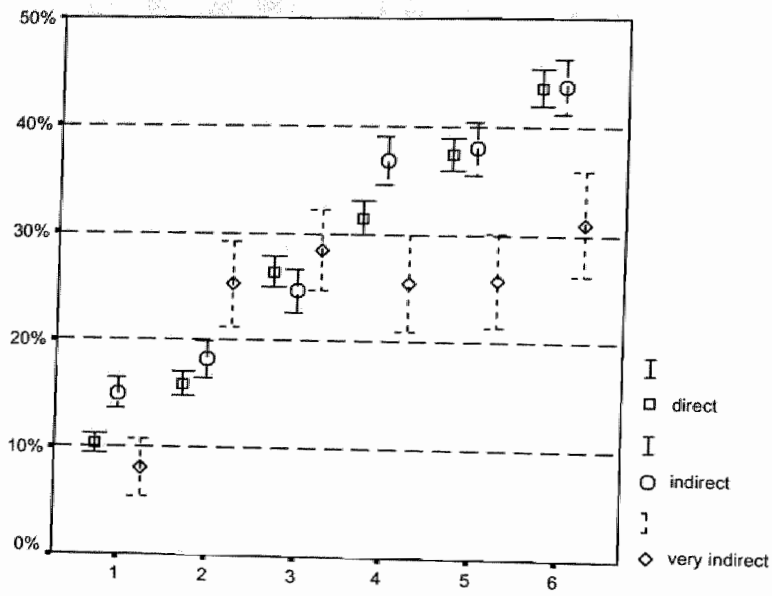
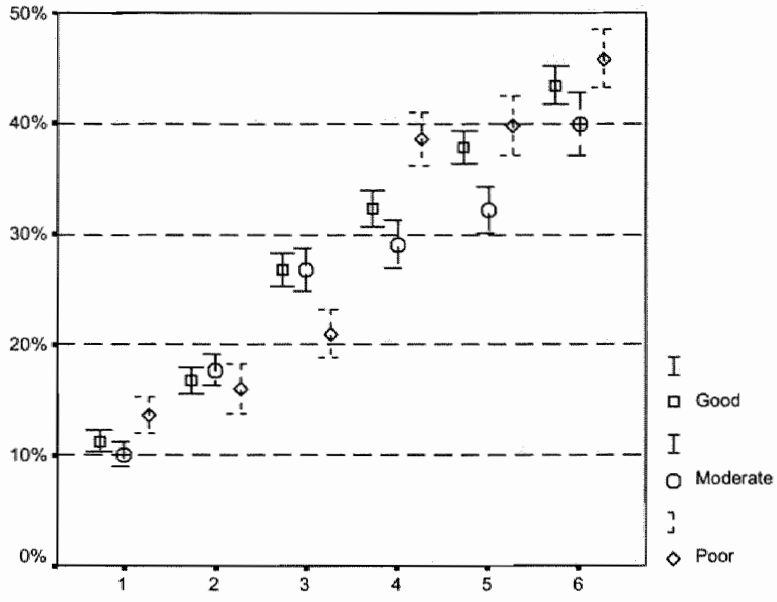


Figure 3b.
The mean percentages scores (correct minus incorrect) and 95% confidence intervals for the three levels of congruence with the Blueprint: good (123 items), moderate (34 items), and poor (22 items).



Discussion

We succeeded in designing a method for determining whether a connection existed between progress test items and the discipline-related objectives in Blueprint 1994. The method proved to be very time-consuming, however. Experienced judges (the authors) needed more than six hours to analyse 230 test items. The associations that were found between test results and the different levels of connection and congruence suggest that a less detailed classification would have sufficed. Only the distinction between connected and unrelated test items yielded significant differences. Very indirectly connected questions can probably better be regarded as unrelated items. This means that steps II and IV of the method could have been omitted. Nevertheless, the approach that was developed to trace a subject in the Blueprint appears to be successful in establishing the presence of a connection, irrespective of whether different types of connection are explicitly distinguished.

What proved problematic was the congruence between levels. We assumed that level T was by definition higher than level D. This may have been too simplistic an assumption and it may not reflect reality. For instance, diagnosing a patient with anaemia would appear to require more knowledge (and skills) than treating the patient. This means that test items

about therapy may be easier to answer than test items about diagnosis. This is in line with what we know from experience: the difficulty of test items is very hard to predict. This is consistent with the discrepancies detected in the discipline-related objectives in this respect. Quite a few objectives are presented with different levels when listed under different disciplines.

Connection and congruence between test items and Blueprint 1994 were better for the clinical sciences items compared with basic sciences and behavioural sciences items. This is not surprising seeing that the latter two categories are not included in the discipline-related objectives. The comparison between test items and Blueprint 1994 in this study was based solely on the discipline-related objectives (clinical pictures and skills). It should be borne in mind that the discipline-related objectives are incomplete for some disciplines, because not all disciplines involved in undergraduate medical education were consulted when the general objectives of Blueprint 1994 were translated into the 'problems as starting points for training' and the discipline-related objectives. It is unlikely, however, that the failure to establish a connection between Blueprint 1994 and test items is attributable solely to omissions in the Blueprint. Part of the explanation may lie in the progress test. The results showed that students' scores on unrelated test items stopped increasing from year 2 onward. By contrast, the scores on connected items continued to increase across cohorts with more years of study. At the end of undergraduate medical education 50% of sixth year students do not answer 31% of the unrelated items compared to 10% of the connected items. Perhaps the curriculum pays no or only limited attention to the unrelated subjects. This suggests that the relevancy of those items may be questionable and that they should not be included in a progress test. On the other hand it was found that a relatively large portion of the unrelated items concerned disciplines which are noticeably (and perhaps unjustifiably) absent from the Blueprint. This confounds the interpretation of the results. The unrelated items may not represent relevant final objectives and the results may reflect a failure of the educational programme.

Finally, we should consider the results from a different perspective. Despite the absence of basic sciences subjects from the Blueprint, it was easy to find a connection with Blueprint 1994 for 73% of the items on those subjects.

If we want to use the progress test to assess students' achievement of the objectives contained in Blueprint 1994, it would seem advisable to encourage the departments to use the Blueprint as a guideline for test item construction. This takes less time than having a committee check the relationship between test items and Blueprint. In Maastricht preparations for such an approach are underway [9].

This article reports on a limited exploratory study in which only one progress test was analysed. As a result the generalizability of the findings is limited. Further analyses of more tests or preferably of all available test data are to be recommended. This study addressed the question as to whether progress test items reflected the objectives stated in Blueprint 1994. We did not ask how many of the subjects included in the Blueprint were covered by

the progress test. In order to answer that question we would have to examine all available progress tests. This would be an interesting albeit time-consuming study. The progress test is only one among many different tests administered in the Maastricht curriculum. Other assessment modalities might be tested for their connection with Blueprint 1994. It is important to remember that medical education pursues other goals besides those that can be assessed by cognitive tests [3]. A sizeable portion of Blueprint 1994 does not allow easy translation into knowledge test items. It is a challenge to medical education to bring the curriculum on paper closer to the learned curriculum and curriculum outcomes [12,13].

References

1 General Medical Council. *Tomorrow's doctors. Recommendations on undergraduate medical education.* London: GMC, 1993.

2 Association of American Medical Colleges. *Physicians for the tweny-first century.* Washington: AAMC, 1984.

3 Metz JCM, Stoelinga GBA, Pels Rijcken - van Erp Taalman Kip EH, Van den Brand - Valkenburg BWM. *Blueprint 1994: Training of doctors in the Netherlands. Objectives of undergraduate medical education.* Nijmegen: University Publication Office, 1994.

4 Hamdy H. Measurements in medical education: Implications for quality [PhD dissertation Rijksuniversiteit Groningen]. Groningen, 2002.

5 Remmen R. An evaluation of clinical skills training at the medical school of the university of Antwerpen [PhD dissertation University of Antwerpen]. Antwerpen, 1999.

6 Raghoebar-Krieger HM, Bender W, Kreeftenberg HG, Stewart RE, Sleijfer DT. Medical students' experiences of diseases in internal medicine in university and community hospitals. *Medical Teacher* 2002; **24**: 402-7.

7 Raghoebar-Krieger HM, Bender W. A comparison of the Dutch blueprint standards (theory) with the experiences of students in clerkships in Groningen (practice). *Journal of Cancer Education* 1997; **12**: 85-8.

8 Van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher* 1996; **18**: 103-9.

9 Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Schuwirth LWT, Van der Vleuten CPM. Quality assurance in test construction: The approach of a multidisciplinary central test committee. *Education for Health* 1998; **12**: 49-60.

10 Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Van der Vleuten CPM. Growth of medical knowledge. *Medical Education* 2002; **36**: 711-7.

11 Van Hessen PA, Verwijnen GM, Imbos TJ. De kennis van de nederlandse basisartsen gemeten met de Maastrichtse voortgangstoets [Knowledge of the Dutch basic physician as assessed with the Maastricht progress test]. *Nederlands Tijdschrift voor Geneeskunde* 1991; **135**: 1975-8.

12 Kassebaum DG, Eaglen RH, Cutler ER. The objectives of medical education: Reflections in the accreditation looking glass. *Academic Medicine* 1997; **72**: 648-56.

13 Harden RM. Developments in outcome-based education. *Medical Teacher* 2002; **24**: 117-20.

Chapter 5

Growth of medical knowledge

Verhoeven BH, Verwijnen GM, Scherpbier AJJA, van der Vleuten CPM
Published in *Medical Education* 2002; 36: 711-7

Introduction

Many studies on clinical reasoning affirm that knowledge is a central factor in medical competence. Medical expertise appears to be based upon doctors' well-developed, highly structured and reshapeable knowledge networks.[1-6] It is therefore important that medical school provides students with a comprehensive and functional knowledge base. To evaluate the effectiveness of the structure and content of medical curricula in this respect, we need to know how students' knowledge grows and develops in the course of their medical training. Much is known about knowledge increment during specific curricular elements or courses, but little is known about the development of knowledge over the medical curriculum as a whole. The latter is the focus of this article.

Several published studies have addressed the growth of medical knowledge during the entire curriculum. However, some are limited to the growth of knowledge in a single discipline or a cluster of disciplines, and most are based on cross-sectional data.[7-9] Others describe the introduction of longitudinal assessment instruments and focus on the reliability, validity, and educational implications of such tests.[10-14] Several publications present mathematical models to predict growth of knowledge.[15-17] To date, no studies have been published about the relationship between content and structure of the undergraduate curriculum and the development of students' medical knowledge base during the entire training programme.

The study reported in this article was performed at Maastricht Medical School, the Netherlands, which claims to offer a problem-based, student-centred, horizontally and vertically integrated curriculum, organised by themes. Students enter the six-year programme directly from secondary education. The first four years consist of mostly six-week, interdisciplinary, thematic units. During years 5 and 6, the clinical phase, students rotate through the major clinical disciplines.[18]

Since 1976, a distinctive feature of the assessment programme has been the progress test (PT), which is administered four times a year to all students regardless of their class. Each PT is a comprehensive examination and it is constructed with the intention to reflect the final objectives of the curriculum.[13] To ensure that tests are equivalent, a blueprint is used, derived from the International Classification of Diseases. In the blueprint, each discipline is assigned to one of three clusters, namely basic sciences (anatomy, biochemistry, pharmacology, physiology, genetics & cell biology, immunology, microbiology, and pathology), clinical sciences (surgery, cardiology, dermatology, obstetrics and gynaecology, family medicine, internal medicine, paediatrics, ENT, neurology, orthopaedics, ophthalmology, pulmonology, radiology, rehabilitation medicine, and urology) and behavioural/social sciences (health care economics, epidemiology, health care law, ethics and philosophy, medical psychology, medical sociology, and psychiatry).[19] The PT contains approximately 250 items in the (multiple) true / false / I do not know format. Content and wording of all items are critically reviewed by a test review committee.[19] Students are discouraged to blindly guess by calculating the test score using formula

scoring. They have the possibility to state they do not know the answer. This is penalized nor rewarded. A correct answer is rewarded with one mark while an incorrect answer is penalized with a negative mark. The sum of the marks is calculated and expressed on a percentage scale. Over the course of the six-year curriculum students sit 24 PTs. This means that the test scores provide cross-sectional and longitudinal data. Students' collective successive PT scores reflect the development of medical knowledge throughout the curriculum, and thus provide an excellent longitudinal, cross-sectional design to study knowledge growth. This paper presents the first exploratory results.

Given the problem-based, student-centred, integrated curriculum, we hypothesised that medical knowledge would show a steady increase throughout the curriculum. After all, encouragement of continuous learning is a major objective of problem-based learning.[20] Assuming an approximately equal curriculum load across years, we expected total PT scores to show a linear upward growth curve. Secondly, the integrative nature of the curriculum should lead to similar continuous growth rates for the three sciences clusters at least during the first four years. The fact that all students spend the last two years rotating through clinical disciplines as clerks, might impede sustained contributions by all sciences clusters during the last phase of the medical training.

Methods

Instrument
To obtain the most reliable estimate possible of the average medical knowledge growth curve throughout the curriculum, we used the scores of all PTs (84) administered between September 1977 and May 1998.

Subjects
All students that entered Maastricht medical school between September 1977 and September 1997 were included in the study.

Procedure
Ideally, undergraduate medical students sit 24 PTs, 16 tests in the first four preclinical years and 8 tests in the two clerkship years, i.e. there are 24 measurement points.[21] Each individual PT score is a measurement of one student's knowledge and represents a dot on his/her knowledge curve across the 24 measurement points. We collected all available individual test scores and used these to calculate the average test score for each of the 24 measurement points. By doing this, the material comprises of groups of individuals of several cohorts, each measured at standardized times. This is called a mixed longitudinal design, which is considered the best design for measuring change.[22] The average test scores across the 24 measurement points were analysed using a curve estimation procedure that calculates the mathematical function that best explains the data (SPSS release 7.5, Nov 14 1996). Both linear ($Y = b_0 + b_1 x$) and quadratic ($Y = b_0 + b_1 x + b_2 x^2$) models were used. Next the "curve of best fit" was plotted, resulting in an estimated growth curve of the

"average" student's knowledge. This procedure was repeated for the subscores on basic, clinical, and behavioural/social sciences.

Table 1.
The number of students that took the PT per measurement moment.

Measurement number	Number of test scores	Measurement number	Number of test scores
1	2996	13	2672
2	3128	14	2656
3	3136	15	2596
4	3109	16	2459
5	2875	17	2191
6	2805	18	2085
7	2792	19	2112
8	2761	20	2071
9	2691	21	1992
10	2679	22	1967
11	2674	23	1797
12	2548	24	1699

Results

The number of students varied considerably across the measurement moments (table 1). Partly this is a result of attrition. A number of students dropped out at various points in the curriculum. This is less than 10%.[23] Most of the variance in the number of subjects is explained by fluctuation of class-size. The Maastricht medical school started in 1974. In 1977 only four classes were present to make the PT. In 1977 no scores were available from measurement moment 17 on forward and in 1978 from measurement 20 on forward. These classes consisted of no more than 60 students. Between 1977 and 1997 the number of students per class has risen drastically each year to 200 in 1997. Moreover, students who have been held back from progression (partly) repeat a year. This way significantly more test scores were available of the first measurement moments. In all, we collected 60491 test scores of 3226 different students. Figure 1a presents the mean total test scores for the 24 measurement points.

Figure 1a.
Mean total PT score per measurement moment.

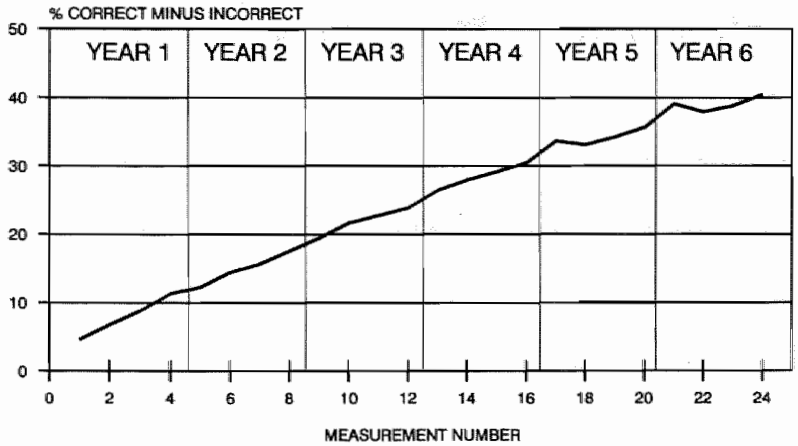
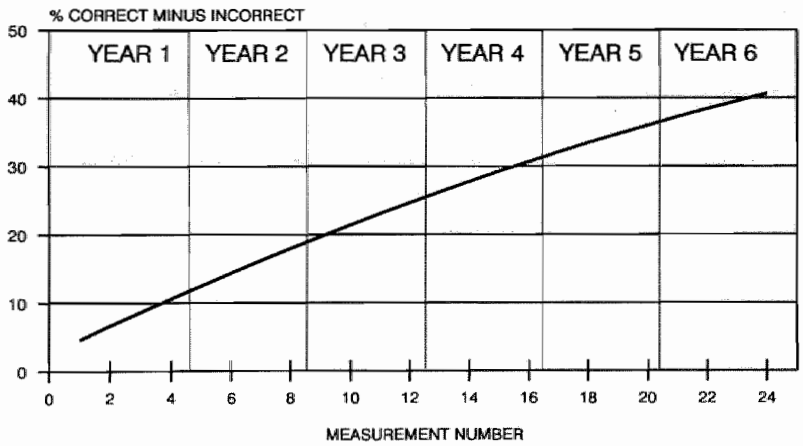
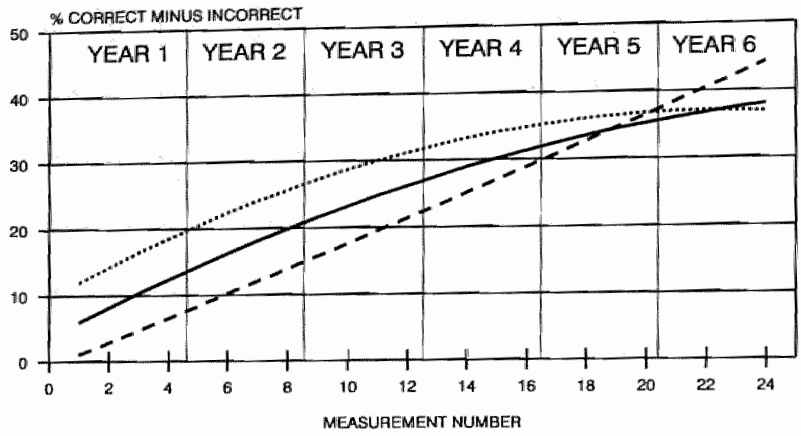


Figure 1b.
The "curve of best fit" which explains the data of figure 1a best.



The line in figure 1b represents the mathematical function ("curve of best fit") that best explains the data. Figure 2 shows the "curves of best fit" for basic, clinical, and behavioural/social sciences scores.

Figure 2.
"Curves of best fit" for basic (solid line), clinical (dashed line) and behavioural/social sciences (dotted line).



Tables 2 and 3 show the calculated best fitting linear functions and quadratic functions of the four curves including R^2 (proportion of explained variance) as a measure of fitting. The growth curves based on the total test score and the scores on basic and behavioural/social sciences are best approached by a quadratic function (i.e. curved). The growth of clinical sciences knowledge can equally well be approached by a linear and by a quadratic model (R^2=0.995). The b_2 in the quadratic model equals 0.0027, meaning it is almost a straight line.

Table 2.
Mathematical equations of the best fitting linear models per growth curve and the proportion explained variance (R^2).

	$Y = b_0 + b_1 x$		
Growth Curve	b_0	b_1	R^2
Total	4.8481	1.5665	0.989
Basic Sciences	7.7091	1.3930	0.960
Social Sciences	15.5639	1.0814	0.884
Clinical Sciences	-1.0760	1.8777	0.995

Table 3.
Mathematical equations of the best fitting quadratic models per growth curve and the proportion explained variance (R^2).

	$Y = b_0 + b_1 x + b_2 x^2$			
Growth Curve	b_0	b_1	b_2	R^2
Total	2.5302	2.1014	-0.0214	0.997
Basic Sciences	3.6191	2.3369	-0.0378	0.987
Social Sciences	9.5559	2.4678	-0.0555	0.973
Clinical Sciences	-0.7853	1.8107	0.0027	0.995

The mean correct minus incorrect score for overall knowledge increases from 5% to 41% during the curriculum. The growth patterns of the three subscores are considerably different. On entering medical school students know the most about behavioural/social sciences and least about clinical sciences. During the first three years basic, behavioural/social, and clinical sciences have similar, nearly parallel growth curves. In year 4 the curve for behavioural/social sciences starts to level off and the same happens to the curve for the basic sciences in year 5. This results in two curve intersections in year 5, because clinical knowledge inclines faster than knowledge of basic and behavioural/social sciences. After six years, the mean score for behavioural/social sciences has increased from 12% to 37%. This contrasts sharply with the scores on basic sciences (from 6% to 38%) and clinical sciences (from 1% to 44%).

Discussion

At the end of the curriculum, medical students score 41% (correct minus incorrect) of the maximum PT score. At this measurement moment students leave 25% of the questions

unanswered and answer 17% of questions incorrectly (data not shown). The correct minus incorrect score of 41% corresponds to a correct score of 58%, which seems rather low for graduating students. However, national and international comparative studies using the PT as measuring instrument reported similar scores.[7,24-26] The average graduate score is significantly higher than the content-based standard established by an expert panel (correct score of 41.4%).[27] The low correct score is partly an artefact of formula scoring. Forcing students to answer all items will, by pure chance alone, result in an increase in the correct scores of some 12.5%. Theoretically, this will result in a correct score of 71%.[28] In the PT, as in all other end-objective tests that are administered to all students regardless of their year of training, an "I do not know" option is inevitable as not all students are expected to have mastered all the objectives included in the test. Graduates do not answer 25% of the questions. Assuming this means "I do not know because the subject was not covered in my years of training", it suggests that the test items are too difficult, controversial or do not reflect the end-objectives of undergraduate medical education. If this is the case, (part of) the PT fails to assess the core content of the curriculum. Further research is needed to test this hypothesis.

Another possibility is that the curricular objectives for years 5 and 6 are not met. The results show that the growth in overall medical knowledge during the undergraduate medical programme is continuous but not quite linear; at the end of the educational programme growth diminishes. The growth in clinical sciences knowledge remains linear. The other two clusters demonstrate slowing growth rates from years four and five. At the end of the medical curriculum, clinical sciences knowledge has increased by 43%, whereas the growth in knowledge of behavioural/social and basic sciences is much less impressive, at 25% and 32%, respectively. These findings are not consistent with the integrative nature of the Maastricht educational approach. A closer look at the actual curriculum is required to explain the findings. The main focus of the interdisciplinary units in years 1 and 2 is on the normal functioning of the human body, and in years 3 and 4 the focus is on abnormal functioning. Finally, all students spend two years in clerkship rotations through eleven clinical disciplines, including family medicine. During the first three years the knowledge curves of the three sciences clusters are very similar, but from year four students appear to pay less attention to basic and behavioural/social sciences. This seems to favour the attention clinical disciplines get.

It could be argued that being a doctor is about being a good clinician, who needs clinical knowledge rather than knowledge of basic or behavioural sciences. However, several studies have shown that clinical expertise is based upon a large body of knowledge from all kinds of disciplines, linked and organised in a knowledge network.[6] This network is shaped by adding associated findings, incorporating pathophysiological knowledge and modification of erroneous parts. Research in cognitive psychology has linked the development of a well-structured associative network of functional knowledge with learning in a professionally meaningful context, with abundant opportunities to practice and apply earlier acquired knowledge.[6] A curriculum that is designed to meet these learning requirements should ideally result in continuous learning in all three broad domains of

medicine throughout the curriculum (and even thereafter). Working on patient problems and trying to find pathological explanations, as students do in the Maastricht curriculum, should lead to a better functioning network, resulting in higher PT scores on all three domains. The marked shift in knowledge growth characteristics at the transition from year four to year five, demonstrated by our data, should not occur. We suspect that this is attributable to the structure of the actual curriculum. Currently, the curriculum is being revised. The sharp distinction between preclinical and clinical years will disappear; patient contacts and clerkships will start much earlier in the educational programme. The patient is to become the basic organising principle, which runs through the curriculum from beginning to end.[29] Normal and abnormal function will be offered together in interdisciplinary units. The assessment programme will be adjusted accordingly. This will encourage students to pay attention to basic, behavioural/social and clinical sciences during each phase of the curriculum.

Methodological considerations

One methodological drawback of this study should be addressed. We used all test scores of all students that had entered our medical school. Although most students do finally graduate, the results may be biased by the scores of students who dropped out at some point in the curriculum. Most students drop out in the first four years or decide not to start their clerkships. Assuming that it is the weaker student who drop out, we conclude that disregarding the scores of the drop-outs would result in slightly higher scores in the first four to five years. This will not change the shapes of the curves essentially. Because more than 90% of the students graduate, we decided to use all available data and accept some possible confounding.[30,31]

Conclusion

In conclusion, the results show that overall knowledge increases monotonously as a function of training time. Growth patterns vary among different discipline clusters. Basic and social/behavioural sciences growth curves level off at senior years of training, which appears to reflect the basic structure of the actual curriculum. Comparative studies of medical schools with different curricula using the same instrument could substantially contribute to our understanding of the growth of medical knowledge and the influence of curriculum characteristics. Several studies of this nature have already been undertaken. Although these were experiments at a single point in time; they clearly showed (international) readiness to collaborate. In 1999, three Dutch medical schools decided to produce and administer the PT together. The initiative contributes to quality assurance in medical education at a national level and has economic benefits at the same time. The next effort will be to develop continuous collaborative research on the growth of medical knowledge in the different curricula and the effects of curriculum changes.

References

1 Neufeld VR, Norman GR, Feightner JW, Barrows HS. Clinical problem-solving by medical students: A cross-sectional and longitudinal analysis. *Medical Education* 1981; **15**: 315-22.

2 Waldrop MM. The necessity of knowledge. *Science* 1984; **223**: 1279-82.

3 Norman GR, Tugwell P, Feightner JW, Muzzin LJ, Jacoby LL. Knowledge and clinical problem-solving. *Medical Education* 1985; **19**: 344-56.

4 Bordage G. Elaborated knowledge: A key to successful diagnostic thinking. *Academic Medicine* 1994; **69**: 883-5.

5 Regehr G, Norman GR. Issues in cognitive psychology: Implications for professional education. *Academic Medicine* 1996; **71**: 988-1001.

6 Van de Wiel MWJ. Knowledge encapsulation. Studies on the development of medical expertise [PhD dissertation Maastricht University]. Wageningen: Ponsen & Looijen, 1997.

7 Verwijnen M, Van der Vleuten C, Imbos T. A comparison of an innovative medical school with traditional schools: An analysis in the cognitive domain. In: Nooman ZM, Schmidt HG, Ezzat ES, editors. *Innovation in medical education: An evaluation of its present status*. Vol. 13. New-York: Springer Publishing Company; 1990: 40-9.

8 Glew RH, Ripkey DR, Swanson DB. Relationship between student's performances on the NBME comprehensive basic science examination and the USMLE step 1: A longitudinal investigation at one school. *Academic Medicine* 1997; **72**: 1097-102.

9 Vosti KL, Bloch DA, Jacobs CD. The relationship of clinical knowledge to months of clinical training among medical students. *Academic Medicine* 1997; **72**: 305-7.

10 Willoughby TL, Hutcheson SJ. Edumetric validity of the quarterly profile examination. *Educational and Psychological Measurement* 1978; **38**: 1057-61.

11 Blake JM, Norman GR, Kinsey E, Smith M. Report card from McMaster: Student evaluation at a problem-based medical school. *Lancet* 1995; **345**: 899-902.

12 Blake JM, Norman GR, Keane DR, Mueller CB, Cunnington J, Didyk N. Introducing progress testing in McMaster university's problem-based medical curriculum: Psychometric properties and effect on learning. *Academic Medicine* 1996; **71**: 1002-7.

13 Van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher* 1996; **18**: 103-9.

14 Boshuizen HPA, Van der Vleuten CPM, Schmidt HG, Machiels-Bongaerts M. Measuring knowledge and clinical reasoning skills in a problem-based curriculum. *Medical Education* 1997; **31**: 115-21.

15 Albers W, Does RJMM, Imbos T, Janssen MPE. A stochastic growth model applied to repeated tets of academic knowledge. *Psychometrika* 1989; **54**: 451-66.

16 Tan ES, Imbos T, Does RJMM. A distribution-free approach for comparing growth of knowledge. *Journal of Educational Measurement* 1994; **31**: 51-65.

17 Tan ES, Imbos T, Does RJMM, Theunissen M. An optimal, unbiased classification rule for mastery testing based on longitudinal data. *Educational and Psychological Measurement* 1995; **55**: 595-612.

18 Van der Vleuten CPM, Scherpbier AJJA, Wijnen WHFW, Snellen HAM. Flexibility in learning: A case report on problem-based learning. *International Higher Education* 1996; **2**: 17-24.

19 Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Schuwirth LWT, Van der Vleuten CPM. Quality assurance in test construction: The approach of a multidisciplinary central test committee. *Education for Health* 1998; **12**: 49-60.

20 Barrows HS, Tamblyn RM. *Problem -based learning. An approach to medical education*. New-York: Springer Publishing Company, 1980.
21 Verwijnen M, Imbos T, Snellen H, Stalenhoef B, Pollemans M, Van Luyk S, Sprooten M, Van Leeuwen Y, Van der Vleuten C. The evaluation system at the medical school of Maastricht. *Assessment and Evaluation in Higher Education* 1982; **7**: 225-44.
22 Van 't Hof MA, Roede MJ, Kowalski CJ. A mixed longitudinal data analysis model. *Human Biology* 1977; **49**: 165-79.
23 Wijnen WHFW. Maastrichts onderwijs en studierendement [Maastricht education and attrition]. In: Nijhof WJ, Warries E, editors. *De opbrengst van onderwijs en opleiding [The output of education and training]*. Lisse: Swets & Zeitlinger; 1986: 165.
24 Van Hessen PAW, Verwijnen GM. Does problem-based learning provide other knowledge? In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, editors. *Teaching and Assessing Clinical Competence*. Groningen; 1990: 446-51.
25 Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Holdrinet RSG, Oeseburg B, Bulte JA, Van der Vleuten CPM. An analysis of progress test results of PBL and non-PBL students. *Medical Teacher* 1998; **20**: 310-6.
26 Albano MG, Cavallo F, Hoogenboom R, Magni F, Majoor G, Manenti F, Schuwirth L, Stiegler I, Van der Vleuten C. An international comparison of knowledge levels of medical students: The Maastricht progress test. *Medical Education* 1996; **30**: 239-45.
27 Verhoeven BH, Van der Steeg AFW, Scherpbier AJJA, Muijtjens AMM, Verwijnen GM, Van der Vleuten CPM. Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. *Medical Education* 1999; **33**: 832-7.
28 Muijtjens AMM, Van Mameren H, Hoogenboom RJI, Evers JLH, Van der Vleuten CPM. The effect of a 'don't know' option on test scores: Number-right and formula scoring compared. *Medical Education* 1999; **33**: 267-75.
29 Scherpbier AJJA, Verwijnen GM, Schaper N, Dunselman GAJ, Van der Vleuten CPM. Vaardigheidsonderwijs nu en in de toekomst [Current and future skills training]. *Tijdschrift voor Medisch Onderwijs* 2000; **19**: 6-15.
30 Onderwijsinstituut Geneeskunde [Institute for Medical Education]. *Managementrapportage onderwijs geneeskunde 2000/2001 [Management report medical education 2000/2001]*. Maastricht: Maastricht University; 2002.
31 Vereniging van Samenwerkende Nederlandse Universiteiten [Association of Universities in the Netherlands]. *Onderwijsvisitatie geneeskunde en gezondheidswetenschappen [Educational review of the faculties of medicine and health sciences]*. Utrecht: VSNU, 1997.

Chapter 6

The effect on reliability of adding a separate written assessment component to an OSCE

Verhoeven BH, Hamers JGHC, Scherpbier AJJA, Hoogenboom RJI, Van der Vleuten CPM
Published in *Medical Education* 2000; **34**: 525-9

Introduction

Since the introduction of the Objective Structured Clinical Examination (OSCE) by Harden and Gleeson, its psychometric properties, including its reliability, have been extensively investigated.[1-3] A major factor affecting the reliability of an OSCE is the so-called case-specificity problem, i.e. the variability in candidates' performances across stations. As a result an OSCE must consist of a large number of stations to obtain reliable scores, which generally means that many hours of testing time are needed.[4-7] Thus, the testing time required from a reliability perspective usually exceeds the resources available in most medical schools. To solve this feasibility problem, several strategies have been suggested. Swanson and Norcini proposed that attention should be focused on the reliability of the decision rather than on the reliability of the scores.[8] In their study they investigated the reliability coefficients associated with different pass-fail cut-off scores. They found striking differences in reliability and testing time, i.e. the greater the distance between the cut-off score and the mean score, the higher the reliability of the decision.[8] This implies that focusing on the decision offers the possibility of achieving adequate reliability with fewer stations. In a study where good correlations between a knowledge test of skills (KTS) and an OSCE were found, Van der Vleuten et al. concluded that a KTS is potentially able to predict scores on a performance-based test.[9] A KTS is relatively cheap and easier to administer compared to an OSCE. Its reliability is known to be good, with less testing time being required. However, simply replacing OSCEs by written tests would have an undesirable effect on students' learning behaviour.[9]

A combination of a written test and an OSCE could bypass this undesirable effect. Replacing one or more OSCE stations by a written KTS component might save resources and increase overall test reliability. It could offer an adequate compromise between the demands of reliability and feasibility. This study investigates the impact on the reliability of test scores of adding a separate written test component, i.e. a knowledge test of skills (KTS), to an OSCE.

Methods

Subjects

From a total class of 204 medical students in their third preclinical year of a six-year program, 38 volunteers were recruited to take a written test in addition to their regular end-of-year OSCE. The students received a small financial compensation for their cooperation. The average OSCE score of these 38 volunteers did not differ significantly from that of the other third-year students ($t=0.622$; $df=202$; $p=0.96$).

Instruments

Written test
A knowledge test of skills (KTS) was developed consisting of 86 true/false questions. This is the test format that is commonly used in our school. An 'I don't know' option was available for each item. The questions were based on the clinical skills standards used in the skills training program of the Maastricht medical curriculum.[10] Some items included pictorial information. The 86 questions covered all the domains (including physical examination, laboratory and communication skills) that a third-year student in the Maastricht medical curriculum is expected to have mastered. The test was administered using a computer. Students viewed the items on screen and their responses were scored immediately. To discourage students from guessing, incorrect answers were subtracted from correct answers. Examples of items are given in figure 1.

Figure 1.
Items from the knowledge test of skills

	true	false	?
In applying a mitella the tail closest to the chest is placed across one shoulder. This is the shoulder at the side of:			
1. the afflicted arm or hand	□	□	□
For the examination of the collateral ligaments of the knee, the knee should be in a certain position. This is:			
2. 30 degrees flexion	□	□	□
An adequate inspection of the retina by means of direct funduscopy requires administration of specific eyedrops prior to inspection. The effect of these eyedrops is:			
3. mydriasis	□	□	□

OSCE
The so-called Skills Test (ST) that students take in their third year was used. The ST consists of eight 15-minute stations. The ST has no further written components or follow-up stations, but is entirely performance-based. Students' performances are scored using detailed checklists tailored to the content of the station. For security reasons, a different test form is used every three hours, with most stations being replaced by other ones equivalent in content according to the test blueprint. To test the entire class, three parallel test forms are used requiring a total of 24 stations.

Procedure
After completing the ST the 38 students were accompanied to a computer room where the KTS was administered under supervision. The students were instructed to complete the test as if it were a regular test, even though the KTS had no influence on pass/fail decisions on the OSCE.

Statistical analysis
For each station the percentage of correct responses was calculated for each student. To ensure equal weighting of both test formats, which was needed for further statistical analysis, the item scores of the KTS were transformed onto a similar scale (correct=100%; don't know=0%; incorrect=-100%). To estimate the reliability of the ST and the KTS separately, generalizability coefficients were calculated for both test formats. The combined reliability of the two test formats and their relative contributions to the overall reliability were estimated using multivariate generalizability theory.[11] In multivariate generalizability theory multiple true (universe) and error scores (variances) can be estimated for each of the components of the measurement and these can be combined into a single composite reliability estimate. Technically, this involves a crossed person-by-item (p x i) design nested within a fixed facet (i.e. test format). This will reveal the relative contributions of each component, which can then be used to optimise the overall composite reliability by varying the sample size of the items and/or the assigned weights within each component (i.e. test format). Multivariate generalizability analysis requires that, within each test component, person and item variance components are estimated as well as a person covariance component. In this study this was complicated by the use of multiple parallel test forms in the ST. Therefore, variance and covariance components were estimated for each test form and averaged across parallel test forms, weighted by the different sample sizes of students. The analysis was repeated for a number of (hypothetical) scenarios, which varied in the number of items/stations and in the weights assigned to the subtests.

Results

Table 1 presents descriptive statistical information, the estimated generalizability coefficients for the two test methods and the observed correlation between the scores of the 38 students obtained with the two methods.

Table 1.
Descriptive statistics and estimated generalizability coefficients for skills test and knowledge test of skills undertaken by 38 students.

	Skills test (ST)		Knowledge test of skills (ST)
Number of items	8		86
Mean score	84.64		52.66
Standard deviation	5.94		10.84
Generalizability coefficient	0.55		0.50
Correlation		0.65	

In table 2 the composite generalizability coefficients are reported for various scenarios with different weights and item/station samples. It was found that all students completed the KTS within one hour. To estimate the composite testing times, it was assumed that 90 KTS items would require one hour of testing time. Scenario A represents the real study sample size of 8 stations and 86 items. In this scenario the weight assigned to individual station scores is the same as that assigned to individual KTS items. The separate weights of the two subtests are determined by the number of stations and items, respectively. This is a highly unrealistic scenario, since it involves a very dominant weighting of the written test with only a small influence of the ST on the overall score. As a result, the estimated composite generalizability coefficient of 0.53 is only marginally higher than the original reliability of 0.50 of the KTS only (table 1). Scenario B involves equal weighting for the two entire tests, which results in a substantial increase in reliability. With this scenario, the combination of methods yields a reliability of 0.68 for about three hours of testing time. For an appropriate appreciation of this increased reliability, a comparison with the expected subtest reliabilities for the extended testing time is fair. This is shown in the last two columns of table 2. With three hours of testing time the ST (12 stations) and the KTS (270 items) would yield reliabilities of 0.65 and 0.76, respectively. Apparently, combining the two methods has some added value with respect to reliability when compared to the reliability of the ST alone. However, when resources are saved by reducing the number of stations in the ST (scenarios C and D), a combination of the methods has a substantial incremental value. Compared with scenario B, the combined reliability decreases slightly, but it is still substantially higher compared to the ST reliability at the projected total testing time (10 and 8 stations for scenarios C and D, respectively).
An option to compensate for the loss of reliability when fewer stations are used would be to increase the number of items in the written test. This was done in scenarios E and F, where the number of KTS items is increased to 120. Scenario F shows that even when the number of stations is halved (4), the loss of reliability is fully compensated (as compared to scenario B) by lengthening the KTS.
One might argue that equal weighting of both test formats overemphasizes the cognitive aspect of clinical skills. To investigate the influence of heavier weighting of the ST,

Table 2.
Estimated generalizability coefficients for combined scores of Skills Test and Knowledge Test of Skills as a function of different subtest weights, number of items/stations and total testing time.

Scenario	ST weight	KTS weight	Number of stations (estimated G-coefficient)	Number of items (estimated G-coefficient)	Approximate total testing time, h:m	Estimated composite G-coefficient	ST reliability at projected total testing time (no. of stations)	KTS reliability at projected total testing time (no. of items)
A	n stations[1]	n items[1]	8 (0.55)	86 (0.50)	3	0.53	0.65 (12)	0.76 (270)
B	50%	50%	8 (0.55)	86 (0.50)	3	0.68	0.65 (12)	0.76 (270)
C	50%	50%	6 (0.48)	86 (0.50)	2:30	0.67	0.60 (10)	0.72 (225)
D	50%	50%	4 (0.38)	86 (0.50)	2	0.64	0.55 (8)	0.68 (180)
E	50%	50%	6 (0.48)	120 (0.58)	2:50	0.72	0.63 (11)	0.75 (255)
F	50%	50%	4 (0.38)	120 (0.58)	2:20	0.68	0.58 (9)	0.71 (210)
G	67%	33%	8 (0.55)	86 (0.50)	3	0.71	0.65 (12)	0.76 (270)
H	67%	33%	6 (0.48)	86 (0.50)	2:30	0.68	0.60 (10)	0.72 (225)
I	67%	33%	4 (0.38)	86 (0.50)	2	0.62	0.55 (8)	0.68 (180)
J	67%	33%	6 (0.48)	120 (0.58)	2:50	0.70	0.63 (11)	0.75 (255)
K	67%	33%	4 (0.38)	120 (0.58)	2:20	0.64	0.58 (9)	0.71 (210)

[1]The weight is directly defined by the number of stations (8) and the number of items (86) in the subtest

scenarios G through K represent replications of scenarios B through F with a two-thirds/one-third weight distribution between ST and KTS. The comparison of scenario G with scenario B, using the sample sizes of the real test used in this study, shows that a small increase in reliability is achieved when the performance-based component is weighted more heavily. With the unequal weight distribution, any reduction in the number of stations leads to slightly lower reliabilities compared to the corresponding equal weighting situation. Increasing the length of the written test to 120 items (scenario K) does not fully compensate for the loss in reliability.

Discussion

This study investigated the impact on reliability of the addition of a separate written test component to a performance-based OSCE. The drawback of this study is its small sample size. Because of the experimental nature of the KTS we were dependent on volunteers, which resulted in only a limited number of students being tested. It would have been more appropriate to randomise. However, the average OSCE score of these 38 volunteers did not differ significantly from that of the other third-year students. The small number of students may also have led to inaccurate estimation of the generalizability coefficients. The values reported in table 2, however, appear to be quite normal. The generalizability coefficient for the results of the real ST in this study is entirely comparable to reliabilities of OSCEs involving the same number of stations that are reported in studies using larger data sets.[12] The same holds for the written true/false test.[13] The correlation between the performance-based component and the written test is also similar to earlier findings. Despite these commonalities, replication in a larger data set would be desirable to provide a firmer basis for our conclusions.

The data showed that a substantial increase in reliability can be achieved by adding a separate written test to an OSCE. Of course, the increase is partly due to the longer testing time needed for the composite test. However, it could be demonstrated that the estimated increased reliability of the composite test exceeds that of an OSCE with the same increased testing time. When the composite reliability is compared to that of a KTS taking the same amount of time, the KTS has a higher reliability. However, replacing an OSCE by a KTS would resolve one problem by creating another one in terms of validity and educational impact. The gain in test reliability would be offset by a loss in educational benefit. Tests do not only assess students' achievements, they also guide students' learning. A paper and pencil knowledge test will overemphasize the cognitive aspects of clinical skills and students will tend to spend less time practising skills if the test does not require them to actually demonstrate these skills.[14] What is needed is a balance between educational advantages, reliable testing, and feasible test procedures.[15] Our data suggest that this can be achieved by shortening the performance-based test component, i.e. using fewer OSCE stations, and compensating for the resulting loss in reliability by adding a separate written test. The conclusion appears justified that in this way a reduction in the resources required can be achieved without compromising reliability. It is important that the weighting of the

test components should also be considered. Obviously, the more weight is placed on the performance-based component, the more difficult it will be to compensate the reduced number of stations by a written component.

It should be noted that our study was performed from the perspective of reliability. With any choice regarding the addition or removal of stations and items, validity, educational and financial issues are involved. These issues may vary from situation to situation, depending on the circumstances. However, from a reliability perspective, adding a separate written component to a performance-based test appears to be a fruitful strategy.

References

1 Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education* 1979; **13**: 41-54.

2 Van der Vleuten CPM, Swanson DB. Practical strategies for improving the reproducibility of tests involving standardized patients. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, editors. *Teaching and Assessing Clinical Competence.* Groningen; 1990: 344-51.

3 Vu NV, Barrows HS. Use of standardized patients in clinical assessment: Recent developments and measurement findings. *Educational Researcher* 1994; **23**: 23-30.

4 Norman G, Feightner J, Tugwell P, Muzzin L, Gyuatt G. The generalizability of measures of clinical problem solving. *Proceedings of the Annual Conference on Research in Medical Education* 1983; **22**: 110-4.

5 Newble DI, Swanson DB. Psychometric characteristics of the OSCE. *Medical Education* 1988; **22**:

6 Swanson DB. A measurement framework for performance-based tests. In: Hart IR, Harden RM, editors. *Further developments in assessing clinical competence*. Montreal: Heal Publications; 1987: 13-45.

7 Van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine* 1990; **2**: 58-76.

8 Swanson DB, Norcini JJ. Factors influencing reproducibility of tests using standardized patients. *Teaching and Learning in Medicine* 1989; **1**: 158-66.

9 Van der Vleuten CPM, Van Luyk SJ, Beckers HJM. A written test as an alternative to performance testing. *Medical Education* 1989; **23**: 97-107.

10 Scherpbier AJJA. Kwaliteit van vaardigheidsonderwijs gemeten [Assessing the quality of skills training] [PhD dissertation Maastricht University]. Maastricht: Datawyse, 1997.

11 Brennan RL. *Elements of generalizability theory*. Iowa City: ACT Publications, 1983.

12 Van der Vleuten CPM, Van Luijk SJ, Swanson DB. Reliability (generalizability) of the maastricht skills test. *Proceedings of the Annual Conference on Research in Medical Education* 1988; **27**: 228-33.

13 Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. *Medical Education* 1985; **19**: 238-47.

14 Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Medical Education* 1983; **17**: 165-71.

15 Van der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education* 1996; **1**: 41-67.

Chapter 7

Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges

Verhoeven BH, Van der Steeg AFW, Scherpbier AJJA, Muijtjens AMM, Verwijnen GM, Van der Vleuten CPM
Published in *Medical Education* 1999; **33**: 832-7

Introduction

Tests and examinations drive student learning.[1,2] For the students, the examination programme is the real curriculum.[3,4] A close match between educational objectives and assessment programme can prevent students from following a "hidden curriculum" of undesirable assessment-based objectives. To achieve this, longitudinal final objective assessment methods have been developed in various medical schools, such as the Quarterly Profile Examination (QPE), the Progress Test (PT) and the Personal Profile Index (PPI).[5-8] These comprehensive examinations reflect the final objectives of the curriculum and sample the complete domain of knowledge that is considered pertinent to undergraduate medical education. These tests are administered periodically (e.g. 3 or 4 times per year) to all medical students regardless of their year of training. This format is intended to reinforce desirable learning behaviour in that it precludes test-directed studying, discourages students from leaving their individual learning paths, encourages functional long-term knowledge and provides feedback to which learning activities can be tailored. Research has shown that these educational objectives are generally attained.[6,7,9,10] Making pass/fail decisions, however, i.e. setting standards, has not yet been addressed properly. Because of the comprehensive nature of progress testing and because each test administration requires a different passing score for each class, setting a passing score is quite complicated. A convenient and widely used standard setting method is a normative approach, with relative cut-off scores being defined by the overall performance of each class. The advantage is that variations in test difficulty are automatically corrected for, but there are also several drawbacks. Firstly, a number of students will always fail, regardless of examinees' abilities. Secondly, the heterogeneity of the student population reduces the validity of the reference group. Thirdly, examinees can deliberately influence the passing score, and, finally, the standard is not known in advance.[11,12] An absolute standard does not have these shortcomings. Its use is appropriate when mastery of content is involved and the percentage of qualified examinees is unknown.[13] This study investigates the use of an Angoff procedure for setting standards for the progress test.[14] The progress test is one of the main instruments used in the Maastricht medical school problem-based curriculum to assess knowledge and reinforce students' self-directed learning. Four progress tests are administered annually to all students (approximately 1000), regardless of their class. The progress test consists of approximately 250 true/false items of different taxonomic levels. It samples knowledge across all disciplines and content areas relevant for the medical degree. The items may include facts and figures or they may contain clinical problem vignettes.[8]

Compared with other types of tests, a progress test poses two additional problems in applying an Angoff procedure. Firstly, an Angoff procedure requires the use of expert judges familiar with students' level of performance.[11,13] Usually, the judges are teachers who are experts in the subject matter being tested.[15] With progress testing, it is difficult to find credible experts for all topics to be tested and experts' familiarity with test-takers' expected levels of performance is questionable, particularly in a problem-based curriculum. We would argue that it are the students who are the only real experts, since they are the "consumers" of the curriculum. Students are also able to conceptualise the target

candidates. Therefore, we decided to use a panel of recently graduated students in an Angoff procedure for setting standards.

The second problem concerns the definition of the "borderline" student, because borderline performance in a progress test depends on how far a student has progressed through the six-year curriculum. The progress test is administered to all classes four times per year, i.e. in the course of the curriculum students are assessed at 24 moments, requiring 24 different standards. The progress test was introduced at Maastricht in 1976, and since that time a vast amount of data has been collected. The data shows that the score follows a specific growth pattern across the 24 measurement points. This implies that when a standard is set for one point in the curriculum, the standards for the other 24 points can be derived mathematically.[8] Since the progress test items should reflect the final objectives of the curriculum, it was decided to estimate the standard for graduation level and define borderline performance in relation to this point in the curriculum (i.e. the time of graduation).

The purpose of this study is to assess the reliability and credibility of the Angoff procedure applied to a sample of items from one progress test, using recently graduated students as judges of borderline performance. Generalizability theory was used to assess reliability. To judge credibility, we examined whether the standard resulting from the Angoff procedure yielded different pass/fail rates compared with pass/fail rates resulting from both a relative and a fixed standard. In addition, we investigated the association of the Angoff item estimates with corresponding item difficulties.

Methods

Materials
Prior to administration a sample of 150 items was drawn from the total of 256 items of the progress test of May 1997. Between administration and calculation of student results, 7 items were excluded from the test due to the routine quality assurance procedure.[8] Of the study sample, 146 items remained.

Judges
The eight participating experts were medical doctors who had graduated from Maastricht University 5 months before the study was conducted (range 1 - 10 months). They were selected on the basis of graduation date, willingness to participate and availability. Two of them were seeking employment, three were residents and three were working on their PhD dissertation. When in medical school, the selected experts had average PT scores, indicating that they did not differ significantly from the rest of their class.

Angoff procedure
To estimate the progress test's passing score based on item content and difficulty, a (modified) Angoff procedure was used.[14,16] The judges were instructed not to apply a

correction for guessing. They were asked to estimate for each item the probability of an imaginary borderline test-taker, at the time of graduation, knowing the correct answer. The judges were given the correct answers. However, they were not given the percentage of examinees that answered each item correctly (p-values) to avoid bias that might affect the assessment of the credibility of the Angoff procedure. Prior to the standard setting procedure, the judges received a letter describing the purpose of the study, the Angoff method and instructions for the day on which the study would take place. The panel of judges was scheduled to meet from 9.00 to 17.00 hours with two breaks, one in the morning and one in the afternoon. The meeting started with a plenary discussion moderated by one of the researchers to establish a working definition of "the borderline student". The following definition was formulated and used during the experiment: "A borderline student is a student who spends an average amount of time studying, whose knowledge is just sufficient to pass at graduate level, but who frequently has difficulty in scoring above the cut-off score of the progress tests." Subsequently, the judges received a booklet containing the selected PT items, each item with the answer key and two blank spaces where the experts could enter their estimates. They were asked to read one item and estimate the percentage of the borderline group that would know the correct answer at the moment of graduation. All estimates were entered in the booklets and written on a whiteboard. The judges with the highest and lowest estimates explained their positions, usually followed by a short debate within the panel. All judges were free to enter an adjusted estimate in the second blank space. In this way, the first 10 items were judged. The judges then made preliminary estimates for the next 10 items and the leader polled the group for their estimates and wrote these on the whiteboard. Whenever there was a discrepancy of 20 or more percentage points between any two judges, a discussion followed. Before the next item was dealt with, the judges were given the opportunity to change their estimate, whether or not a debate had taken place. This procedure was repeated for all the remaining items.

Statistical analysis

Descriptive statistics

To judge the representativeness of the item selection, scores on sampled and non-sampled items were compared. Each judge's Angoff estimates were averaged across all items to establish the passing score per judge. Mean and standard deviations were calculated. All judges' estimates (converted to a percentage scale) on all individual items (8 x 146) were averaged to establish the passing score for the test.

Reliability

Generalizability theory was used to investigate the reliability.[17-19] As all items were rated by all judges, a crossed item-by-judge design ANOVA (i x j) was used, followed by variance component estimation using the GENOVA package.[20] Since we wish to estimate the error of setting a standard for a given test, the variance of the item main effect was not included in the error variance. The Root Mean Square Error (RMSE) was estimated and expressed accordingly:

$$RMSE = \sqrt{\frac{\hat{\sigma}_j^2}{n_j} + \frac{\hat{\sigma}_{ij}^2}{n_i n_j}}$$

Where n_i is the number of items, n_j is the number of judges, and $\hat{\sigma}^2$ is the estimated variance component for the associated effect. The RMSE is an estimate of the standard error of the mean of Angoff estimates across items and judges. It indicates the error involved in the test's passing score.[18]

Credibility

Because the judges were asked not to apply a correction for guessing, the Angoff estimate can best be compared to the correct score. For the item sample a percentage correct score was calculated for each of the judges and used to determine a pass or a fail for the sixth year students only. The pass/fail rate obtained by using the Angoff standard was compared with the pass/fail rates obtained with a relative standard (mean test score of the sixth year students minus one standard deviation) and a fixed standard, respectively. The fixed standard was derived from the average pass/fail rates obtained with a relative standard for past test performances across 8 years.[12] Furthermore, the Pearson correlation was calculated for the mean of the judges estimates per item and the p-values (percentage correct).

Results

Table 1 presents the mean test score and standard deviation of the 69 sixth year medical students that took the progress test of May 1997. Descriptive statistics are presented for the total progress test and the two subtests (sampled and non-sampled part). In addition, the reliabilities of the three tests are given.

Table 1.
Mean progress test scores of sixth year medical students (n = 69) and reliabilities of the total progress test, set of selected and non-selected items for the Angoff procedure.

	Number of questions	Mean correct score (%)	SD	Reliability*	Standardized reliability†
Total PT	249	53.0	8.7	0.90	0.90
Sampled	146	52.3	8.4	0.83	0.89
Not sampled	103	53.8	10.2	0.83	0.92

* Cronbach's alpha † Reliability corrected for the reduced number of items by using the Spearman Brown prophecy formula; standardization is towards 249 items.[21]

The average student results on the total PT and the two sub-tests are comparable. These scores are also comparable to those of previous progress tests on the same occasion. There seems to be a common, relatively stable score that students achieve at the end of the undergraduate medical curriculum. For a correct interpretation of the scores, one should take into account that students can use a question mark option (zero points) and that students are instructed that incorrect answers will be penalized with minus 1 (-1) point. The true correlation (corrected for attenuation) between the sampled items and the total PT is 1.00 (observed correlation: 0.96).[22] The sample used for this study appears to be representative. The mean Angoff estimate (i.e. the average estimation of all items and judges) is 41.4% with a standard deviation of 1.7. Table 2 presents the descriptive statistics of the Angoff estimates for each judge separately.

Table 2.
Mean and standard deviation of the Angoff estimates (146 items).

Judge	1	2	3	4	5	6	7	8
Mean (%)	41.1	41.3	42.7	40.1	42.3	38.3	41.6	43.7
SD	22.8	20.5	22.3	24.3	20.6	22.6	22.6	22.3

The mean total ratings of the 8 judges show little variation. However, the standard deviations per judge are quite large, implying that the Angoff estimates differ considerably across items.

Table 3 presents the results of the analysis of variance and the estimated variance components.

Table 3.
Analysis of variance and estimated variance components.

Source of variability (effect)	d.f. *	Sum of squares for score effects (SS)	Mean squares (MS)	Estimated variance component	SE †	Percentage of total variance
Items	145	486013.26	3351.82	407.89	48.87	81.8
Judges	7	2868.27	409.75	2.20	1.32	0.4
IJ, e	1015	90065.73	88.73	88.73	3.94	17.8

* degrees of freedom
† standard error of estimated variance component

Approximately 82% of all variance can be attributed to variation between items. Apparently, a wide range of item difficulties is found in the progress test. In line with results from table 2, the percentage of variance associated with consistent variability of judges across items is very small (0.4% of the total estimated variance). Even the overall error term is relatively small with approximately 18% of the total variance.

Table 4 reports the Root Mean Square Errors of the test's passing score as a function of the number of judges and the number of items in the test.

Table 4.
Root Mean Squared Errors as a function of the number of items and the number of judges.

	Number of Judges									
Number of Items	3	4	5	6	7	8	9	10	11	12
50	1.15	1.00	0.89	0.81	0.75	0.70	0.66	0.63	0.60	0.58
100	1.01	0.88	0.79	0.72	0.66	0.62	0.59	0.56	0.53	0.51
150	0.96	0.84	0.75	0.68	0.63	0.59	0.56	0.53	0.50	0.48
200	0.94	0.81	0.73	0.66	0.61	0.57	0.54	0.51	0.49	0.47
250	0.92	0.80	0.71	0.65	0.60	0.56	0.53	0.51	0.48	0.46
300	0.91	0.79	0.71	0.64	0.60	0.56	0.53	0.50	0.48	0.46

The RMSE is the error of the test's passing score expressed on the original scoring scale (i.e. the percentage correct scale). With 8 judges and 150 items, an RMSE of 0.59 was achieved. An approximately double (i.e. 1.96) RMSE yields a 95% confidence interval for the test's passing score (41.4% ± 1.2%). This interval is relatively small compared to the standard deviation of the test scores (8.7%). However, with normally distributed test scores, a 1% shift of the passing score changes the failure rate by approximately 2.5%. This implies that we should aim at a precision of at least 1% on the scoring scale, which corresponds to an RMSE of 0.51. This would be achieved with 10 judges, each rating 200 items or more.

Table 5 shows the percentages of sixth year students failing the progress test using the standard arrived at by the Angoff estimate and the failure rates when applying a fixed and a relative standard, respectively.

Table 5.
Failure rates of sixth year medical students (n = 69) in the progress pest for different standards.

Standard used	Passing score (%)	Failure rate (%)
Angoff standard	41.4	7.2
Fixed standard	52.4	55.1
Relative standard	43.9	10.1

The failure rate obtained by the Angoff standard is lowest and comes closest to the failure rate of the relative standard. Finally, the correlation between item-difficulties (p-values) and item Angoff estimates was 0.83, indicating that the Angoff estimate does include item difficulty variation.

Discussion

The large variation in item difficulties is typical for progress tests. Although the test is targeted to the final objectives of the curriculum and thus should yield high scores on all items at the final administration, this is routinely not the case and a considerable spread of item scores is found. Repeated and intensive feedback to item authors about scoring profiles of individual items and total tests, and specific instructions and workshops about item construction and test design have failed to bring about a reduction in the spread of item scores. Apparently, this is an inherent phenomenon of progress tests. More than 80%

of the variance could be attributed to item variance, leaving only a small portion of judge and other error variance. The small size of the judge variance is probably also partly a result of the procedure followed. The judges were allowed to revise their estimates, i.e. the individual estimates are not fully independent.

In the estimation of the error involved in the Angoff estimate across judges and items (the test's passing score), the item difficulty variance is not considered as part of the error because the students will be tested on this set of items and generalization is not towards other items. Only the (small) judge variance and overall error term (including the interaction effect between judges and items) are considered as error variance. Adding judges would considerably improve the reproducibility of the passing score. With 10 judges in the panel judging 200 items or more, an acceptable precision is reached, i.e. 1% on the scoring scale. In other words, with the normal sample size of items (approximately 250) an Angoff panel should consist of 10 judges.

Credibility was assessed by comparing the student failure rate associated with the Angoff method with those of two conventional standards, and by investigating the relationship between the Angoff score and item difficulty. Although different standards yield different outcomes, some credibility of the (modified) Angoff procedure can be inferred from the comparison in this study. The Angoff standard came closest to the relative standard, based on the mean and the standard deviation of this progress test. Unlike the fixed standard, the relative standard takes account of variations in test difficulty. Compared to the fixed standard (52.4%), the relative standard (43.9%) is lower, indicating that the difficulty of this test is above average. This is also supported by the same finding in the other five year groups that took the test on the same occasion. The scores obtained on previous tests by the group of sixth year students give no reason to expect this group to differ from other groups on this occasion in the curriculum. The necessity to include test difficulty into the judgment of a passing score is evident given the magnitude of item variance involved in progress testing. Credibility was also supported by the high correlation of the item Angoff estimates with the actual item scores. It shows that the Angoff standard is sensitive to item and thus to test difficulty.

In conclusion, the results of this study suggest that the Angoff procedure is an appropriate standard setting method for a progress test. The use of recently graduated students as judges appears to be justifiable. Feasibility could be a problem, since considerable resources are required for reaching a reproducible passing score estimation. Further research could focus on the effect of minimizing the resources and logistics needed for the Angoff procedure. For instance, one might look at the use of different groups of judges judging fewer items per test or explore the reduction of the panel size by providing initial estimates by item authors and/or reviewers.

References

1 Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Medical Education* 1983; **17**: 165-71.

2 Frederiksen N. The real test bias. Influences of testing on teaching and learning. *American Psychologist* 1984; **39**: 193-202.

3 Van der Vleuten C, Newble D, Case S, Holsgrove G, McCann B, McRae C, Saunders N. Methods of assessment in certification. In: Newble D, Jolly B, Wakeford R, editors. *The certification and recertification of doctors*. Cambridge: Cambridge University Press; 1994: 105-25.

4 Van der Vleuten CPM. Beyond intuition [Inaugural lecture Maastricht University]. Maastricht: Datawyse, 1996.

5 Arnold L, Willoughby TL. The quarterly profile examination. *Academic Medicine* 1990; **65**: 515-6.

6 Blake JM, Norman GR, Kinsey E, Smith M. Report card from McMaster: Student evaluation at a problem-based medical school. *Lancet* 1995; **345**: 899-902.

7 Blake JM, Norman GR, Keane DR, Mueller CB, Cunnington J, Didyk N. Introducing progress testing in McMaster university's problem-based medical curriculum: Psychometric properties and effect on learning. *Academic Medicine* 1996; **71**: 1002-7.

8 Van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher* 1996; **18**: 103-9.

9 Van Berkel HJM, Nuy HJP, Geerlings T. The influence of progress test and block tests on study behaviour. *Instructional Science* 1995; **22**: 317-33.

10 Van Til CT. Voortgang in voortgangstoetsing. Studies naar de aansluiting van de voortgangstoets op probleemgestuurd onderwijs [Progress in progress testing. Studies on the suitability of progress testing within a problem-based educational context] [PhD dissertation Maastricht University]. Wageningen: Ponsen & Looijen, 1998.

11 Norcini JJ, Shea JA. The credibility and comparability of standards. *Applied Measurement in Education* 1997; **10**: 39-59.

12 Muijtjens AMM, Hoogenboom RJI, Verwijnen GM, Van der Vleuten CPM. Relative or absolute standards in assessing medical knowledge using progress tests. *Advances in Health Sciences Education* 1998; **3**: 81-7.

13 Norcini JJ. Research on standards for professional licensure and certification examinations. *Evaluation and the Health Professions* 1994; **17**: 160-77.

14 Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL, editor. *Educational measurement*. Washington DC: American Council on Education; 1971: 508-600.

15 Impara JC, Plake BS. Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement* 1998; **35**: 69-81.

16 Livingston SA, Zieky MJ. *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service, 1982.

17 Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley & Sons Inc., 1972.

18 Brennan RL. *Elements of generalizability theory*. Iowa City: ACT Publications, 1983.

19 Shavelson RJ, Webb NM, Rowley GL. Generalizability theory. *American Psychologist* 1989; **44**: 922-32.

20 Crick JE, Brennan RL. *A general purpose analysis of variance system [computer program]*. Version 2.2. Iowa City: The American College Testing Program, 1984.

21 Crocker LM, Algina J. Procedures for estimating reliability. *Introduction to classical & modern test theory*. Orlando: Harcourt Brace Jovanovich College Publishers; 1986: 131-56.
22 Nunnally JC. Theory of measurement error. *Psychometric theory*. 2nd ed. New York: Mc Graw-Hill Book Company; 1978: 190-224.

Chapter 8

Panel expertise for an Angoff standard setting procedure in progress testing Item writers compared to recently graduated students

Verhoeven BH, Verwijnen GM, Muijtjens AMM, Scherpbier AJJA, Van der Vleuten CPM
Published in *Medical Education* 2002; **36**: 860-7

Introduction

When test results are used for pass/fail decisions, robust standards are needed. An Angoff procedure can be used to establish an absolute passing score for a test.[1] This procedure involves estimation of the performance of borderline examinees by a panel of judges. In a recent study an Angoff procedure to establish a passing score for the Progress Test (PT) was evaluated.[2] The PT is a longitudinal assessment method and is best characterized as a comprehensive final examination in medicine that is regularly taken by all students. Every PT samples the complete domain of factual knowledge that a medical student must have mastered upon graduation.[3] The panel of judges in an Angoff procedure should be familiar with the performance level of the students who take the test and they should be credible experts in all topics being tested.[4] Usually, these expert judges are teachers. It is assumed that they are able to conceptualise the borderline test-taker and predict their performance on each individual test item.

Previous studies have shown that teachers are generally accurate in ranking items according to difficulty, but inaccurate in estimating actual levels of examinee performance.[5] In a previous Angoff procedure for the PT we used recently graduated students as judges. It was argued that students, being the 'curriculum consumers', are to some extent content experts in the knowledge domain assessed by the PT. In addition, we expected the recently graduated students to be better equipped to conceptualise the target candidates than were the teachers. This argument should particularly hold for a PT, which assesses the level of knowledge to be attained at the end of undergraduate medical training. The results of that study suggested that the use of student judges is justifiable. The recent graduates appeared to give a good estimate of the actual level of examinee performance.[2] There are, however, some drawbacks associated with the use of recent graduates as judges. Firstly, both faculty and the public are likely to question their expertise and objectivity. Secondly, a minimum of ten judges was needed to reach an estimated passing score of acceptable precision.[2] This means that substantial resources are required, which affects the feasibility of this standard setting procedure. Thirdly, the more people are involved in the standard setting procedure, the more difficult it is to maintain secrecy. These drawbacks might be overcome if the panel of judges consists of teachers who are also the item reviewers or item writers. The use of such a panel would enable the Angoff procedure to be incorporated into the regular item review procedure, which would greatly enhance feasibility.[6] The study reported in this article investigated the feasibility of an Angoff procedure using item writers as judges. The main focus of the study was on: 1) the reliability and credibility of the Angoff procedure using item writers as judges when applied to a sample of items from a PT; and 2) a comparison between the results of an Angoff procedure with item writers as judges and an Angoff procedure where the judges were recent graduates.

Methods

Background

The Maastricht undergraduate medical curriculum lasts 6 years. It is problem-based, student-centred and organized in themes. The PT is one of the main instruments used for the assessment of knowledge. It is a written test consisting of approximately 250 items. There are three answer options: true, false and don't know. The test contains items from all relevant disciplines and domains and reflects the level of knowledge which students are expected to have attained at the end of the undergraduate curriculum. Guessing is discouraged by the use of formula scoring in which the number of incorrect scores is subtracted from the number of correct ones and the don't know option yields no marks. Annually, four PTs are administered to all medical students. This amounts to 24 measurement points over the course of the curriculum (6 years x 4 tests). Although test content is the same for all students, the pass/fail standard becomes more strict from year 1 through year 6. A study revealed that the scores of the students show a characteristic and stable growth pattern across the 24 measurement points.[3] This implies that when a standard is set for one measurement point in the curriculum, the standards for the other 23 points can be derived mathematically.[7] Since the PT items reflect the final objectives of the curriculum, it was decided to estimate the standard for graduation level (24th measurement point).

Materials

A sample of 150 items was drawn from the 256 items of the PT of May 1997 prior to administration of the test. The routine quality assurance procedure led to the removal of 7 items after test administration, leaving a study sample of 146 items.[6] This sample was also used in the Angoff procedure conducted earlier with a panel of recent graduates.[2]

Judges

The 14 participating teachers/item writers were selected on the basis of affinity with education, exposure to students, familiarity with the PT, willingness to participate and availability. The teachers' professions, functions and departments are listed in table 1. All teachers were or had been item writers for their discipline. Four of them had been item reviewers in the recent past.

Table 1.
Profession, function and department of participating faculty members and mean and standard deviation of the Angoff estimates (146 items).

Judge	Profession	Function	Department	Mean (%)	SD
1	Family Physician	Teacher / Educationalist	Skillslab	53.2	22.6
2	Rheumatologist	Teacher / Consultant	Internal Medicine	57.0	27.4
3	Pulmonologist	Teacher / Consultant	Internal Medicine	57.6	23.7
4	Sociologist	Teacher / Therapist	Psychiatry	53.4	19.0
5	Psychologist	Teacher / Therapist	Psychology	48.5	23.3
6	Epidemiologist	Teacher / Researcher	Epidemiology	53.4	27.1
7	Biochemist	Teacher / Researcher	Biochemistry	56.9	20.9
8	Pathologist	Teacher / Consultant	Pathology	55.0	25.5
9	Family Physician	Teacher / Physician	Skillslab	57.5	21.6
10	Paediatrician	Teacher / Consultant	Paediatrics	53.3	32.4
11	Physiologist	Teacher / Researcher	Physiology	53.4	21.1
12	Family Physician	Teacher / Physician	Family Medicine	46.6	25.7
13	Family Physician	Teacher / Physician	Family Medicine	52.0	26.6
14	Family Physician	Teacher / Educationalist	Skillslab	55.6	22.0

Angoff procedure
A (modified) Angoff procedure was used to estimate the passing score of the PT based on item content and difficulty.[1,8] The judges were instructed not to apply a correction for guessing. They were asked to imagine a borderline examinee and estimate for each item the probability that that examinee would know the correct answer. The judges were given the correct answers to the test items but not the actual percentage of correct answers given by the students who took the test (*P*-values). Prior to the standard setting procedure, the judges received a letter describing the purpose of the study, the Angoff method, and instructions for the day when the Angoff procedure would be conducted. The panel of judges met twice, on two separate days, from 9.00 a.m. to 1.00 p.m.. The first meeting started with a plenary discussion moderated by one of the researchers and aimed at establishing a working definition of "the borderline student". The following definition was formulated and used during the experiment: "A borderline student is a student who spends an average amount of time studying, whose knowledge is just sufficient to pass at graduate level, but who frequently has difficulty in scoring above the cut-off score on the progress test." A booklet containing the PT items to be judged was handed out. Each item was provided with its answer key, and two blank spaces where the teachers could enter a first estimate and, if deemed appropriate, a second, adjusted estimate. The judges were asked to read one item and enter their estimate of the percentage of the borderline group that would give the correct answer at the 24th measurement point, i.e. near graduation. The estimates were then written on a whiteboard. After the judges who gave the highest and lowest estimates explained their decisions, the panel debated the estimates briefly. All judges were free to

enter an adjusted estimate in the second blank space. In this way, the first ten items were judged. The judges then made preliminary estimates for the next ten items and these were written on the whiteboard. Discrepancies of 20 or more percentage points between any two judges were discussed and judges could enter adjusted estimates. This procedure was repeated until the judges had given Angoff estimates for all the items in the sample.

Statistical analysis

The estimates entered in the second space, i.e. the adjusted estimates were included in the analysis. If no adjusted estimate was given, the first estimate was used.

Descriptive statistics

To judge whether the item sample was representative of the PT, the test scores on sampled and non-sampled items were compared. The passing score per judge was established by averaging the Angoff estimates of each judge across all items. Means and standard deviations were calculated. The passing score of the test was established by calculating the mean of the estimates of all the judges for all individual items (14 x 146).

Reliability

Generalizability theory was used to investigate the reliability of the standard.[9-11] As all items were rated by all judges, a crossed item-by-judge design ANOVA (i x j) was used, followed by variance component estimation using the GENOVA package.[12] Since we wanted to estimate the error of setting a standard for a specific PT, the variance of the item main effect was not included in the error variance. The Root Mean Squared Error (RMSE) was estimated as follows:

$$RMSE = \sqrt{\frac{\hat{\sigma}_j^2}{n_j} + \frac{\hat{\sigma}_{ij}^2}{n_i n_j}},$$

where n_i is the number of items, n_j is the number of judges, and $\hat{\sigma}^2$ is the estimated variance component for the associated effect. The RMSE is an estimate of the standard error of the mean of the Angoff estimates across items and judges. It represents the variation in the passing score when the same items are judged by a different sample of judges.[10]

Credibility

The judges were asked not to apply a correction for guessing. Assuming that guessing is effectively discouraged by formula scoring, the Angoff estimate can best be assumed to refer to the correct score. The actual test results of each student were used to calculate their percentage correct score for the study sample. On the basis of these scores a pass or fail for the sixth year students was determined according to three different standards: the Angoff standard, the relative standard (mean test score of the sixth year students minus one standard deviation), and a fixed standard based on past test performance across eight years (mean test score on eight tests minus one SD).[7,13] The Pearson correlation was calculated for the mean of the judges' estimates per item and the *P*-values (percentage correct). High correlations can be viewed as an indicator of validity.[14]

Item writers versus graduates
We compared the variance components, RMSE, pass/fail rate and correlations of the Angoff estimate of the item writers with those of the recent graduates from our earlier study. [2].

Results

The PT of May 1997 was taken by 69 sixth year students. The study sample appears to be representative of the total test. The mean score, standard deviation and reliability of the sampled items and those of the non-sampled items are comparable.[2] The correlation of the test scores from the study sample and the total PT was 0.96 (estimated true correlation: 1.00).[15] The mean Angoff estimate (i.e. the average estimation for all items and judges) is 53.8% with a standard deviation of 3.2 across judges. Table 1 presents descriptive statistics of the Angoff estimates by judge. The mean ratings of the fourteen judges showed a range of 11%. The standard deviations per judge were quite large (24% on average). This implies that the Angoff estimates differ considerably across items. Table 2 presents the results of the analysis of variance and the estimated variance components.

Table 2.
Analysis of variance and estimated variance components.

Source of variability (effect)	d.f.*	Sum of squares for score effects (SS)	Mean squares (MS)	Estimated variance component	SE#	Percentage of total variance
Items	145	803816.02	5543.56	380.54	46.19	62.8
Judges	13	20031.57	1540.89	9.07	3.85	1.5
IJ, e†	1885	407210.07	216.03	216.03	7.03	35.7

* degrees of freedom
\# standard error of estimated variance component
† overall error term

Approximately 63% of all variance can be attributed to variation between items. Apparently, a wide range of item difficulties is found in the PT. In line with the results from table 1, the percentage of variance associated with differences between judges that are consistent for all items is small (1.5% of the total variance). The judge-item interaction term is considerable and accounts for approximately 36% of the total variance. Note, however, that with $n_i = 146$ or more, the RMSE is dominated by the main judge effect ($\hat{\sigma}_j^2$), because the judge-item interaction ($\hat{\sigma}_{ji}^2$) is attenuated by factor $1/n_i$:

$$RMSE = \sqrt{\frac{\hat{\sigma}_j^2}{n_j} + \frac{\hat{\sigma}_0^2}{n_i n_j}} = \sqrt{\frac{\hat{\sigma}_j^2}{n_j} + (\hat{\sigma}_{ij}^2 \times \frac{1}{n_i} \times \frac{1}{n_j})}$$

$$RMSE = \sqrt{\frac{9.07}{14} + (216.03 \times \frac{1}{14} \times \frac{1}{146}} = \sqrt{0.648 + 0.106}$$

Table 3 reports the Root Mean Squared Error of the passing score as a function of the number of judges and the number of items in the test.

Table 3.
Root Mean Squared Error as a function of the number of items and the number of judges.

Number of Items	Number of Judges									
	3	5	10	15	20	30	39	40	50	60
50	2.11	1.64	1.16	0.94	0.82	0.67	0.59	0.58	0.52	0.47
100	1.93	1.50	1.06	0.87	0.75	0.61	0.54	0.53	0.47	0.43
150	1.87	1.45	1.03	0.84	0.72	0.59	0.52	0.51	0.46	0.42
200	1.84	1.42	1.01	0.82	0.71	0.58	0.51	0.50	0.45	0.41
250	1.82	1.41	1.00	0.81	0.70	0.58	0.50	0.50	0.45	0.41
300	1.81	1.40	0.99	0.81	0.70	0.57	0.50	0.49	0.44	0.40

With 14 judges and 150 items, a RMSE of 0.87 was achieved. The corresponding 95%-confidence interval for the standard is 53.8% ± 1.71% (the estimated standard ± 1.96 x RMSE). This interval is relatively small compared to the standard deviation of the test scores (8.7%). However, with normally distributed test scores, a 1% shift in the passing score changes the failure rate by approximately 2.5%. We therefore argue that a precision of at least 1% on the scoring scale, which corresponds to a RMSE of 0.51, would be a minimally acceptable benchmark. This would be achieved with 39 judges and 200 items or more.

Table 4 shows the failure rates of sixth year students derived from the standard arrived at by the Angoff estimate, the fixed standard and the relative standard, respectively. The Angoff standard yields the highest failure rate, followed by the fixed standard. The correlation between item difficulty (*P*-values) and item Angoff estimate was 0.82.

Table 4.
Failure rates of 6th year medical students in the Progress Test for different standard setting methods.

Standard	Passing score	Failure rate	
	(%)	n	(%)
Angoff standard (item writer panel)	53.8	39	56.5
Fixed Relative Standard	52.4	38	55.1
Relative Standard	43.9	7	10.1
Angoff standard (graduates panel)	41.4	5	7.2

Item writers versus graduates
On average the panel of item writers expected a borderline score that was 12.4% higher than that expected by the panel of recent graduates. The correlation across items of the Angoff estimates of the two panels is 0.84. This is very high compared to correlations reported in the literature. This may be due in part to the wide range of item difficulties in the PT. Table 5 summarizes the main differences between the two panels.

Table 5.
Summary of the main differences between the two Angoff panels.

	Angoff panel	
	graduates	item writers
Standard (%)	41.4	53.8
Standard deviation across items and judges	1.7	3.2
Correlation p-value – Angoff estimate	0.83	0.82
Variance attributed to item variance (% of total variance)	82	63
Variance attributed to judge variance (% of total variance)	0.4	1.5
Number of judges needed (n)	10	39
Failure rate (%)	7.2	56.5

Discussion

Absolute standard
The PT is used in a formative way to inform students about their progress, possible gaps in their knowledge and their relative position with regard to the end-objectives. The test is also used to reach pass/fail decisions. When test results are used for pass/fail decisions,

robust standards are needed. During the first 20 years a relative standard was used with relative cut-off scores being defined by the overall performance of each class (mean minus standard deviation).[13] The advantage of such an approach is that variations in test difficulty are automatically corrected for. There are also several drawbacks: 1) the standard is not known in advance; 2) a relative standard results in a number of failing students, regardless of examinees' abilities; 3) examinees can deliberately influence the passing score; and 4) heterogeneity of the student population reduces the validity of the standard.[4,7] An absolute standard does not have these shortcomings. Its use is appropriate when mastery of content is involved and the percentage of qualified examinees is unknown.[16] The introduction of a PT re-examination and political pressure made an absolute passing score for the PT necessary.

Reliability

In the current study the judge variance (expressed as a percentage of the total variance) in the Angoff standard is three times as large as the judge variance in the study with recent graduates as judges (1.5% versus 0.4%). The variance attributed to error variance is twice as large as the corresponding variance obtained with the Angoff procedure using recently graduated students as judges (35.7% versus 17.8%).[2] Adding judges would improve the reproducibility of the passing score. However, an acceptable precision of 1% on the scoring scale, would require a panel of 39 item writers judging 200 test items. It would appear that recently graduated students show more agreement and produce more reliable Angoff estimates than a panel of item writers. Several factors could attribute to this difference. The range between the highest and lowest primary estimate was higher in the item writer panel than in the graduate panel. This could be due to the difference in composition of the two panels. The graduate panel consisted only of recently graduated medical students, a very homogenous group. The item writer panel was much more heterogeneous and consisted of physicians, researchers, consultants, therapists and educationalists; all with different educational backgrounds and different experiences with students. After discussion, the graduates adjusted their estimates more than the item writers did after the short debate. Why the graduates were more inclined to compromise is not clear. Maybe peer group effects or their experiences in small group education play a role. Providing the two groups with simultaneous and extensive identical Angoff procedure training might have levelled the variance. The definition of "the borderline student" as we used it in this experiment could lead to more variance in the panel of item writers than in the graduate panel. It is possible that recently graduated students are more able to conceptualise the target candidates than the teachers because all graduates took the PT during six years while in medical school while only few of the item writers made PTs themselves. It is possible (some of) the item writers judge on the basis of what they think students ought to know.

Feasibility

A procedure involving 39 item writers would require enormous resources and be very lengthy. Moreover, it is not possible to conduct a fruitful debate with 39 participants. If the discussions were omitted and individual estimates used, the number of judges needed to

obtain a reliable passing score would increase to 77. It is questionable whether such a large number of suitable item writers could be recruited four times a year to judge 200 items.

Credibility
The high correlation of the item Angoff estimates with the actual item scores shows that the Angoff standard is sensitive to item and thus test difficulty. The necessity to include test difficulty in the judgment of a passing score is evident given the magnitude of item variance involved in progress testing.[7] Both graduates and item writers appear to be quite capable of incorporating item difficulty into their estimates. However, the difference in the test failure rate between the two Angoff procedures is unacceptable. More than 55% of candidates would fail when the standard established by the item writers is used. The standard established by the recent graduates, on the other hand, would yield a lower failure rate than the relative standard.

Scoring methods
In this study the Angoff estimate is compared to the correct score. However, in actual practice, the "don't know" option and formula scoring are used. If formula scoring is effective in making students select the "don't know" option instead of taking a gamble, the correct score is the best indicator of knowledge. However, one might argue that formula scoring will not (fully) eliminate guessing. When some of the correct answers are indeed obtained by guessing, the correct score might spuriously flatter the knowledge level of the students. If there is no "don't know" option, the number of incorrect answers is a good indicator of the number of correct answers obtained by guessing (assuming a student will be correct in 50% of guesses and incorrect in the other 50%). In that situation formula scoring is the most appropriate indicator of knowledge. The "correct-minus-incorrect score" as a measure of knowledge for tests that do include the "don't know" option, will underestimate students' knowledge. The true knowledge level will lie somewhere between these two extremes. If the above is taken into account, the results of the study would be magnified. The failure rates for all standards would rise, resulting in even more unacceptable outcomes for the item writers standard and less lenient outcomes for the recently graduated students standard.

On the whole, item writers and recent graduates are equally capable of ranking the items according to difficulty. However, student performance tends to be overestimated by the teachers/item writers. The latter finding is in agreement with the literature.[5,14,17,18] Providing the judges with information about item difficulty might solve this problem. However, because we want to set the standard of the PT prior to test administration, only *P*-values of test items used in previous tests could be used. Many items change during the quality control cycle.[6] Every change will alter some of the item characteristics and thus test difficulty. Norcini reported that the use of *P*-values in standard setting procedures has an effect on the ranking of items by experts but does not affect the standard itself.[17] Future research is needed to evaluate the possible benefit of such an adjustment.

Conclusion

The observed differences between the two panels are considerable and have a substantial impact on the passing scores and thus on the failure rate. Using item writers as judges to obtain a reliable passing score is not feasible and the established passing score seems less credible. A panel of recently graduated students yields a reliable albeit lenient standard, but the political acceptability of these judges is doubtful. The findings of the study suggest that a mixed panel (item writers and graduates) might achieve a compromise standard. Such a standard might be neither too stringent nor too mild, have an acceptable precision with feasible panel sizes and be politically acceptable. Future research will have to demonstrate whether these expectations are justified.

References

1 Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL, editor. *Educational measurement*. Washington DC: American Council on Education; 1971: 508-600.

2 Verhoeven BH, Van der Steeg AFW, Scherpbier AJJA, Muijtjens AMM, Verwijnen GM, Van der Vleuten CPM. Reliability and credibility of an angoff standard setting procedure in progress testing using recent graduates as judges. *Medical Education* 1999; **33**: 832-7.

3 Van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher* 1996; **18**: 103-9.

4 Norcini JJ, Shea JA. The credibility and comparability of standards. *Applied Measurement in Education* 1997; **10**: 39-59.

5 Impara JC, Plake BS. Teachers' ability to estimate item difficulty: A test of the assumptions in the angoff standard setting method. *Journal of Educational Measurement* 1998; **35**: 69-81.

6 Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Schuwirth LWT, Van der Vleuten CPM. Quality assurance in test construction: The approach of a multidisciplinary central test committee. *Education for Health* 1998; **12**: 49-60.

7 Muijtjens AMM, Hoogenboom RJI, Verwijnen GM, Van der Vleuten CPM. Relative or absolute standards in assessing medical knowledge using progress tests. *Advances in Health Sciences Education* 1998; **3**: 81-7.

8 Livingston SA, Zieky MJ. *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service, 1982.

9 Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley & Sons Inc., 1972.

10 Brennan RL. *Elements of generalizability theory*. Iowa City: ACT Publications, 1983.

11 Shavelson RJ, Webb NM, Rowley GL. Generalizability theory. *American Psychologist* 1989; **44**: 922-32.

12 Crick JE, Brennan RL. *A general purpose analysis of variance system [computer program]*. Version 2.2. Iowa City: The American College Testing Program, 1984.

13 Wijnen WHFW. Onder of boven de maat. Een methode voor het bepalen van de grens voldoende/onvoldoende bij studenten. [Above or below par. A method to determine the pass/fail cutoff point in student assessments] [PhD dissertation Rijksuniversiteit Groningen]. Amsterdam: Swets & Zeitlinger, 1971.

14 Goodwin LD. Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. *Applied Measurement in Education* 1999; **12**: 13-28.
15 Nunnally JC. Theory of measurement error. *Psychometric theory*. 2nd ed. New York: Mc Graw-Hill Book Company; 1978: 190-224.
16 Norcini JJ. Research on standards for professional licensure and certification examinations. *Evaluation and the Health Professions* 1994; **17**: 160-77.
17 Norcini JJ, Shea JA, Kanya DT. The effect of various factors on standard setting. *Journal of Educational Measurement* 1988; **25**: 57-65.
18 Livingston SA, Zieky MJ. A comparative study of standard-setting methods. *Applied Measurement in Education* 1989; **2**: 121-41.

Chapter 9

The consequential validity of the progress test An investigation into the relationship between test results and problem-based learning behaviour

Verhoeven BH, Van Til CT, Verwijnen GM, Scherpbier AJJA, Van der Vleuten CPM
Under editorial review

Introduction

Students have a preferred learning style but can and will adapt their way of learning to what the learning environment requires of them.[1,2] What, when, how and how much students learn is to a large extent directed and shaped by assessment programmes. The way assessment is programmed appears to have a major effect on study strategy.[3-7] Many assessment programmes are structured in such a way that tests are in competition with each other and invite students to peak from hurdle to hurdle.[7,8] Students prepare the expected content of the next test, which leaves little room to follow individual learning paths or gain in depth understanding. Such test programmes reinforce rote memorization (of facts) and short-term learning objectives.[1,9]

Progress test

In 1976 progress testing was introduced at the Maastricht medical school to reinforce self-directed learning behaviour. The progress test (PT) was designed to preclude test-directed studying and discourage students from leaving their individual learning paths.[10-13] The PT is a comprehensive final examination reflecting the (cognitive) end-objectives of the medical curriculum. It samples knowledge across all disciplines and content areas in medicine relevant for the medical degree. This means that there is no direct relationship between the progress test and any specific course. The candidates remain ignorant of the exact content of the test (all topics of medicine can be tested), so students have little incentive for test-directed studying. Conversely, all individual study activities will be rewarded (at least samplewise). Four times per academic year all students (year 1 through 6) take the same PT simultaneously. For each occasion a new test is constructed. For practical reasons (large numbers of items and students) the test items are restricted to a closed format, i.e. the True / False / Question Mark format.[14] PTs contain about 250 items covering fifteen categories derived from the International Classification of Diseases (ICD).[12] Formula scoring is used to discourage guessing. The test score is calculated by subtracting the number of incorrect answers from the number of correct answers. Students can use a question mark option (no credits) to indicate that they do not know the answer. The correct minus incorrect score is expressed on a percentage scale. The test results reflect how far students have progressed toward the final objectives of undergraduate medical education.

The study reported in this article investigated the consequential validity of the PT.[7,15-17] In other words we wanted to find out if the PT does reinforce desired learning behaviour. In a survey students were asked about their learning behaviour. Different types of learning behaviour were quantified and correlated with PT results. To be able to put the results into perspective we also analysed the relationship between learning behaviour and two course related tests.[10] Unlike the PT, these two tests are not intended to preclude test-directed preparation and may encourage students to leave their individual learning paths. We hypothesized that 1) students who show desired problem based learning behaviour score higher on all three tests and low scoring students display less desirable learning behaviour, 2) learning behaviour is a better predictor of the variance in PT scores than of the variance

in the scores on the course-related tests, and 3) students score higher on the course-related tests, but not on the PT, if they study specifically for these tests. For the three tests the following research questions were posed:
1) Do students with desirable learning behaviour score higher on the test?
2) Are low scoring students showing different learning behaviour compared to high scoring students?
3) Does specific preparation for tests result in higher test scores?
4) How much of the variance in test scores can be explained by learning behaviour?

Methods

Participants
All students in the first four (pre-clinical) years of the undergraduate curriculum of Maastricht medical school who attended the tutorial groups in the last but one unit of the academic year 1996-1997 (N=710) were asked to participate. Students in year 5 and 6 were not included in the study, because this is the clerkship phase and students do not attend tutorial groups nor write unit tests.[18]

Instruments
Participants were asked to complete a questionnaire about problem-based learning behaviour and a short questionnaire on test preparation and amount of study time. The results on the progress test and two course related tests were used as outcome variables.

Course related achievement tests

Unit test
Each pre-clinical year of the undergraduate curriculum consists of seven six week courses called units. A unit test (UT) is administered at the end of each unit. The UT consists of knowledge questions that reflect the content of the course.[19] The purpose of the UT is to assess students' knowledge about the content of the unit and to provide students with information about their mastery of the course objectives.[10,19] The UT consists of 140 to 160 knowledge questions of the True / False / Question Mark format. Formula scoring is used and the score is expressed on a percentage scale.

Skills test (OSCE)
Once a year all students have to demonstrate their skills in a skills test. This test is modelled on the Objective Structured Clinical Examination (OSCE).[20] A test consists of a two-hour circuit in which students perform different skills in six to twelve stations. The content of the skills test (further referred to as OSCE) is based on the educational programme of each year. Performance is rated by trained observers on checklists in which the required task is operationally itemized. Station scores are calculated as the percentage of correct scores on the checklist. The overall test score is calculated by averaging the station scores.

Questionnaire

Many views and models of learning styles have been developed and described in the literature. Learning styles are complex and any model is by definition an oversimplification. A well-known classification is based upon the distinction between deep and surface learning.[3,9,21] Students will use different learning styles in different phases of the learning process because behaviour arises from an interaction between personality traits and situations.[21] In small group learning differences between students become manifest in group interactions. The verbal interaction in a group only shows the tip of the iceberg of the cognitive and metacognitive processes of the individuals who are engaging in the interaction..[22] Students are known to be able to be covertly active, i.e. they use certain study activities in their mind, but do not verbalize them in the group. By assessing personal traits in specific situations, a more reliable picture of the real behaviour can be obtained.[23] In this study we collected data about learning behaviour in different phases of the learning process.

Van Til et al. developed a questionnaire about problem based learning behaviour based on two dimensions of learning: style (deep – surface) and activity (active – passive).[24,25] The dimension "activity" is based upon the many ways and intensities students use to activate and elaborate their knowledge.[26,27] The questionnaire consists of 24 vignettes. Each vignette reflects one of the four combinations of "style" and "activity" that characterise students' learning behaviour: 1) deep-active, 2) deep-passive, 3) surface-active, and 4) surface-passive. For each different stage of the problem based learning approach used in the small group sessions in Maastricht (clarify terms and define the problem, analyse the problem and organize potential explanations, formulate learning objectives, individual study, report and discuss newly acquired information) four vignettes were designed. Table 1 shows the vignettes for "formulating student generated learning issues". In addition, four vignettes were formulated describing behaviour during small-group evaluation. Students were asked to indicate in how far they recognised their own behaviour in the vignettes by rating the vignettes on a 5-point Likert scale (1 = not at all to hardly, 2 = a little, 3 = moderately, 4 = to a large extent or 5 = (almost) completely).

Table 1.
The four vignettes for "formulating student generated learning issues".

PBL learning behaviour		style of PBL learning behaviour	
		deep	surface
activity of PBL learning behaviour	active	"Normally, after brainstorm and analysis, I'm able to formulate the topics of the learning objectives. I'm also able to formulate them in a clear and workable manner. I generally report the ideas I have. The learning goals finally formulated by the tutorial group cover the ones I thought were of interest."	"Usually I state my ideas for learning goals. However, I find it difficult to define topics for the learning objectives after the brainstorm and analysis. I also find it difficult to formulate them in a clear and workable manner. My contribution often helps other students to make a more explicit and clear proposal."
	passive	"Particularly, I listen to other students' suggestions for learning objectives. I can see which issues require further exploration. I generally wait for others to come up with suggestions. Eventually, the final learning issues are very similar to the ones I thought were relevant."	"Particularly, I listen to the ideas of the other students. I mostly wait before expressing my ideas. I think about ideas, but for me it's difficult to identify the main topic in a discussion and the most important learning objectives. I also find it difficult to judge the proposals of others on clarity, specificity and relevance."

Test preparation and study time
To assess test driven learning, two questions asked students to indicate on a 5-point Likert scale (1=not at all to hardly; 5=(almost) completely) to what extent they studied specifically for the PT and the UT. Students were also asked to estimate the overall number of hours they spent on study activities. The questionnaire ended by asking students to indicate the number of hours studied for the PT, UT, and OSCE, respectively.

Procedure

After permission was granted by faculty, course coordinators and tutors, the questionnaires were handed out to the students during the tutorial group session in the fifth week of the last but one course of the academic year 1996-1997. Students were informed about the purpose of the study and received verbal and written instructions. They were asked to fill out the questionnaire based upon their learning behaviour during the past academic year and give their university registration number. Three days later, in the next group session, the questionnaires were handed in and the students received a small reward for their time and effort. Non-responders were contacted by telephone and stimulated to hand in the questionnaire.

Analyses

Questionnaire

The questionnaire was divided into four parts each consisting of six vignettes representing one of the four combinations of learning style and activity. The scores on the vignettes were averaged to obtain per student a mean Likert score (between 1 and 5) for each type of learning behaviour, i.e. "deep-active", "deep-passive", "surface-active" and "surface-passive". The mean scores on "deep-active" and "deep-passive" were summed for each student as were the scores on "surface-active" and "surface-passive". By subtracting the latter from the first, a score (between +4 and –4) on the dimension "style" was calculated. The higher this score, the more a student used a deep learning strategy. The level of activity was determined by adding the scores on "deep-active" and "surface-active" and subtracting the summed scores on "deep-passive" and "surface-passive". The more active a student was, the higher the value. Next, each student was assigned to one of four categories of problem based learning behaviour depending on the type of learning behaviour on which the student had obtained the highest score (mean of six vignette scores). If a student had the highest score on two types of learning behaviour, the student was not assigned to a category but labelled as "no classification".

Achievement tests

The test scores on the OSCE, the last two UTs and the last two PTs of the academic year 1996-1997 were collected at the end of the year. The scores of all students in year 1 through 4 (responders and non-responders) were analysed. For each year of study a standard score (z-score) was calculated for each test.[28] The two UT z-scores were averaged. The same was done for the two PT z-scores. When one of the two z-scores was missing, one z-score was used. If for an individual student the OSCE score or both UT scores or both PT scores were missing, the student was not included in that particular analysis.

Statistical analysis
All calculations and statistical procedures were performed with SPSS (release 10.0.5, Nov 27 1999).

Descriptive statistics
Response rate and number of questionnaires that could be analysed were calculated. The average z-scores on the UT, OSCE and PT of the participating students and the non-participating students were compared. The distribution of problem-based learning behaviours was given and for each type the associated mean and standard deviation of the three test scores, hours studied and intensity of test-directed preparation were calculated across students and years. These descriptive statistics were also calculated for style and activity with each combination of learning behaviours. Departure of equality of variance and normality was tested by Levene's homogeneity-of-variance test and the Kolmogorov-Smirnov test. Symmetry was tested by exploring histograms and calculating skewedness and kurtosis. If the assumptions of the parametric tests were met, differences between groups were tested with the student t-test, one-way ANOVA and Friedman two-way ANOVA. For the other data non-parametric tests were used (Mann-Whitney U, Kruskal-Wallis H and Wilcoxon signed rank test for paired data). A $p < 0.05$ was considered significant.
Next the students were divided into a group of high performing students (students who scored above the 50^{th} percentile) and a group of low performing students (students who scored below the 50^{th} percentile). The mean scores on activity and style of the students with high or low performance on the PT, UT and OSCE, respectively were calculated and compared. The frequencies of types of learning behaviour of the high-scoring versus low-scoring students on the three tests were depicted in a cross table. Differences were tested with the Chi-square test. $P < 0.05$ was considered significant.

Correlation and regression
Correlations between style, activity, hours studied, level of test preparation and test scores were calculated using the Pearson correlation coefficient. A $p < 0.01$ was considered significant. Using multiple hierarchical stepwise linear regression analysis the impact of style and activity of PBL learning behaviour on students' achievement was examined.[28] Using the z-scores, the attribution of year of training had no spurious effect on the regression. Test directed preparation and number of hours studied were incorporated in the model. To investigate any differences between years, the regression analyses were repeated for each year of training separately.

Results

Descriptive statistics

Response
Of the 710 students that were asked to participate, 529 (75%) handed in the questionnaire. Nine questionnaires were not used because answers were missing or erroneous or no university registration number had been provided. 73% of the distributed questionnaires were analysed. To determine whether or not respondents were representative of the total student population, the tests scores of respondents and non-respondents were compared. Respondents scored higher on the UT ($t = 3.127$, $p < 0.01$) and PT ($z = 2.482$, $p = 0.01$), but not on the OSCE ($z = 0.923$, $p = 0.36$).

PBL learning behaviour
Table 2 presents the distribution of the four types of learning behaviour. The corresponding means and standard deviations of the calculated values of style and activity are also shown. Of the 520 students, 64 (12%) could not be classified. Students with deep-active learning behaviour had the highest values on the dimensions style and activity, with surface-passive students obtaining the lowest values. Surface-active students scored lower on style than did deep-active and deep-passive students and lower on activity compared to deep-active students. Deep-passive students scored high on style and low on activity, but not as high as the deep-active students on style and not as low as the surface-passive students on activity. Differences between the four types of problem based learning behaviour in style and activity are both significant ($\chi^2 = 255$, $p < 0.001$ and $\chi^2 = 278$, $p < 0.001$ respectively).

Table 2.
The distribution of the types of problem based learning behaviour across all years of study. On the right, the means and standard deviations of the calculated values for style and activity per type are depicted.

PBL learning type	N(%)	style		activity	
		mean	sd	mean	sd
deep active	213(41)	1.71	0.92	2.34	1.08
surface active	113(22)	-0.33	0.79	1.49	1.08
deep passive	102(20)	1.24	1.03	-0.62	0.87
surface passive	28(5)	-1.01	0.71	-1.63	1.29
no classification	64(12)	0.46	0.72	1.11	1.19

Research question 1:
Do students with desirable learning behaviour score higher on the test?
Table 3 reports means and standard deviations of the z-scores on PT, UT and OSCE for each of the types of learning behaviour. The highest scores on all tests were obtained by students with deep-active learning behaviour, whereas the students with surface-passive behaviour obtained the lowest scores. This effect was most pronounced for the PT and least pronounced for the OSCE. The differences were significant for the PT ($F(3, 450) = 14.36$, $p < 0.001$) and UT ($F(3, 450) = 5.14$, $p = 0.002$), but not for the OSCE.

Table 3.
Means and standard deviations of the z-scores on PT, UT and OSCE for each of the PBL learning types.

	PT			UT			OSCE		
PBL learning type	mean	(sd)	n	mean	(sd)	n	mean	(sd)	n
deep-active	0.36	(0.98)	213	0.25	(0.84)	213	0.19	(0.77)	212
surface-active	-0.13	(0.69)	112	-0.03	(0.83)	112	0.06	(0.72)	111
deep-passive	-0.01	(0.84)	101	-0.01	(0.84)	101	-0.07	(1.03)	99
surface-passive	-0.54	(0.84)	28	-0.24	(0.80)	28	-0.26	(0.96)	28
total	0.10	(0.91)	454	0.09	(0.85)	454	0.07	(0.84)	450

Research question 2:
Are low-scoring students showing different learning behaviour compared to high-scoring students?
The frequencies of types of learning behaviour in the low and high scoring groups of students for the different tests are depicted in a cross table (table 4a). In the groups with high scores on the PT and UT the percentage of students with deep-active learning behaviour (65% and 60%, respectively) was significantly higher than in the low scoring groups. Students with surface-passive learning behaviour were found more frequently in the low scoring groups (82% and 71%). No differences were found for surface-active and deep-passive learning behaviour. For the OSCE no differences were found between the high and low scoring students. Table 4b shows the means on activity and style of the two groups for PT, UT and OSCE, respectively. The students with high scores on the PT and UT had a more active and deep learning style. The OSCE showed no differences.

Table 4.
Percentage of low and high scoring students by learning behaviour for PT, UT and OSCE (a). Mean scores on activity and style of high and low scoring students for PT, UT and OSCE (b).

a	percentage of students in low and high scoring group											
	PT				UT				OSCE			
PBL learning behaviour	low	high	X^2	p	low	high	X^2	p	low	high	X^2	p
deep active	35	65	33.8	0.00*	40	60	13.4	0.00*	47	53	1.18	0.30
surface active	54	46	1.1	0.29	54	46	1.1	0.39	53	47	0.6	0.44
deep passive	53	47	0.6	0.44	54	46	1.0	0.32	53	47	0.3	0.56
surface passive	82	18	12.2	0.00*	71	29	5.4	0.02*	54	46	0.2	0.69

b	mean score of students in low and high scoring group											
	PT				UT				OSCE			
dimension	low	high	z	p	low	high	z	p	low	high	z	p
style	0.55	1.22	6.1	0.00*	0.66	1.11	4.6	0.00*	0.84	0.93	1.1	0.29
activity	0.91	1.53	4.4	0.00*	0.93	1.51	3.8	0.00*	1.12	1.33	1.4	0.17

* $p < 0.05$

Research question 3:
Does specific preparation for tests result in higher test scores?
For the UT 53% of students indicated to prepare themselves moderately, to a large extent or (almost) completely while for the PT this percentage was 35%. Table 5 shows the number of hours studied and the hours spent preparing for specific tests. The number of hours of preparation differed significantly between the three tests (Friedman two-way ANOVA; $\chi^2 = 750.7$, $p < 0.001$). Students spent on average 55 hours on the OSCE, 20 hours on the UT and only 3 hours on the PT. In the same table the extent to which the students study for tests is shown. Students indicate that they do more test-directed preparation for the UT (2.7) than for the PT (2.2) (Wilcoxon signed rank test for paired data; $z = 6.92$, $p < 0.001$). Compared to the other students, students with "surface-passive" learning behaviour spent less time on their study ($z = 2,61$, $p = 0.009$) and expressed the strongest agreement that they engaged in test-directed preparation for the UT ($z = 4,86$, $p < 0.001$). No differences were found for the PT and the OSCE in this respect. Compared to the other three groups, students with deep-active learning behaviour spent more hours studying ($z = 2.58$, $p = 0.01$) but spent least time preparing tests ($z = 2.57$, $p = 0.01$). They also score lowest on the items about specific preparation for the PT ($z = 2.20$, $p = 0.03$) and UT ($z = 4.20$, $p < 0.001$).

In table 6 the correlation matrix for learning style and activity, hours studied, test preparation and test scores is presented. The total number of hours studied correlated positively with UT and OSCE scores, but not with PT scores. The more hours students spent on specific preparation for the OSCE, the higher their OSCE scores. The hours spent on specific preparation for the PT or UT are not correlated with test scores. Specific UT

preparation is negatively correlated with all three test scores while specific PT preparation has no relationship with any of the scores. A greater percentage of time spent on test preparation is negatively correlated with number of hours of total study time and PBL learning behaviour style.

Table 5.
The hours studied, the hours of test directed study, and the extent to which students claim to study for the tests.

	total	deep active	surface active	deep passive	surface passive	Kruskal-Wallis H	
N	520	213	113	102	28		
	mean (sd)	mean (sd)	mean (sd)	mean (sd)	mean (sd)	X^2	p
hours studied	28 (9)	29 (9)	27 (8)	27 (10)	24 (8)	11.46	0.010*
hours studied for test							
unit test	20 (20)	18 (18)	21 (15)	19 (18)	23 (16)	12.00	0.007*
progress test	3 (4)	3 (4)	2 (3)	3 (5)	3 (5)	4.22	0.238
OSCE	55 (40)	59 (47)	54 (33)	53 (32)	43 (29)	2.18	0.536
% time spent on tests	20 (14)	17 (11)	21 (12)	19 (15)	27 (23)	11.24	0.010*
explicit study for UT†	2.7 (1.3)	2.4 (1.3)	3 (1.3)	2.6 (1.2)	3.9 (1.0)	38.06	0.000*
explicit study for PT†	2.2 (1.3)	2 (1.2)	2.4 (1.4)	2.1 (1.4)	2.6 (1.6)	8.27	0.041*

† mean on a 5-point Likert scale, 1 = not at all to hardly, 5 = (almost) completely
* $p < 0.05$

Table 6.
The correlation matrix for PBL-style and –activity, hours studied, number of hours of specific test preparation, the level of test preparation and test scores.

	test-score			% TP#
	PT	UT	OSCE	
PBL style	0.30*	0.20*	0.07	-0.21*
PBL activity	0.19*	0.14*	0.17*	-0.05
study-hours				
total	0.11	0.29*	0.14*	-0.20*
PT	-0.05	-0.04	0.10	0.21*
UT	-0.10	0.02	0.01	0.77*
OSCE	0.10	0.11	0.19*	0.36*
PT preparation	-0.07	-0.01	0.07	0.18*
UT preparation	-0.26*	-0.27*	-0.18*	0.28*

* $p < 0.001$
# %TP = percentage of total study hours spent on test preparation

Research question 4:
How much of the variance in test scores can be explained by learning behaviour?
Style and activity were found to be significantly correlated with PT and UT scores. There was also a positive correlation between activity and OSCE scores (table 6). The results of the regression analyses that examined the impact of learning behaviour on PT, UT and OSCE scores are shown in table 7. The results indicate that only a small part of the total variance is explained by the variables used in this model, from 6.4% for the OSCE to 13.8% for the UT. However, the increase in explained variance that occurs when the different variables are added one by one, showed that specific test preparation and number of hours studied contribute most to the variance in UT and OSCE scores (adjusted R^2 = 11% and 4% respectively) and least to the variance in PT scores (adjusted R^2 = 2%). For learning behaviour the reverse was found. Style and activity explained 10% of the variance in PT scores and only 2% of the variance in UT and OSCE scores. The data for each year separately revealed that problem based learning behaviour contributes significantly to the total explained variance in year 2 and 3. In the fourth year 30% of the variance in PT scores could be explained by style and activity. For the UT and OSCE no significant increase in explained variance occurred when problem based learning behaviour was added to the regression model.

Table 7.
The results of the multiple hierarchical stepwise linear regression analysis used to examine the impact of hours of study, test preparation and PBL learning behaviour on the PT, UT and OSCE scores.

	R^2	B	seB	beta	t	p	R^2 change	F	p
PT									
1. hours of study	0.013	0.008	0.005	0.081	1.805	0.072	0.013	6.282	0.013*
2. test preparation	0.023						0.009	2.288	0.103
explicit study for test		-0.007	0.035	-0.010	-0.194	0.847			
hours of study for test		-0.012	0.012	-0.051	-0.945	0.345			
3. PBL learning behaviour	0.122						0.099	26.452	0.000*
style		0.203	0.033	0.287	6.209	0.000*			
activity		0.046	0.024	0.085	1.870	0.062			
(constant)		-0.344	0.144		-2.391				
UT									
1. hours of study	0.080	0.021	0.004	0.221	4.746	0.000*	0.080	41.827	0.000*
2. test preparation	0.116						0.036	9.598	0.000*
explicit study for test		-0.088	0.031	-0.139	-2.867	0.004*			
hours of study for test		0.000	0.002	0.000	-0.006	0.995			
3. PBL learning behaviour	0.138						0.022	6.213	0.002*
style		0.084	0.032	0.128	2.646	0.008*			
activity		0.034	0.023	0.067	1.500	0.134			
(constant)		-0.358	0.169		-2.115				
OSCE									
1. hours of study	0.026	0.009	0.005	0.097	1.895	0.059	0.026	11.384	0.000*
2. test preparation	0.044						0.017	7.662	0.006*
hours of study for test		0.003	0.001	0.129	2.526	0.012*			
3. PBL learning behaviour	0.064						0.021	4.677	0.010*
style		-0.010	0.032	-0.015	-0.308	0.759			
activity		0.074	0.025	0.150	3.009	0.003*			
(constant)		-0.384	0.128		-2.997				

* $p < 0.05$

Discussion

This study focused on individual differences between students in learning behaviour in a problem-based learning context and the relationships between learning behaviour and achievement on different kinds of tests. In general, test results were favourably affected by active and deep learning behaviour. Style appeared to contribute more to PT and UT scores whereas activity contributed more to the OSCE score. Of the three tests studied, the PT

showed the lowest correlation between results and test-directed preparation and the highest correlation with desired learning behaviour. These results are in line with those of earlier studies.[25,29,30]

Learning behaviour
Students who reported deeper and more active learning behaviour scored higher on the PT. Low scoring students reported more passive and surface learning behaviour. Learning behaviour explained 10% of the variance in PT scores and only 2% of the variance in UT and OSCE scores. The amount of explained variance was limited. This may be attributable to the homogeneity of the study population in respect of learning behaviour. Only 2% of students were categorized as "surface-passive" whereas 41% fell into the category "deep-active". Although 75% of the overall student population completed the questionnaire, there was an overrepresentation of high achievers. This is in line with Hilliard's findings.[2] It is possible that a higher response rate would have increased the number of students who use surface learning approaches, although Van Til found the same distribution with a response rate of 94%.[25] It is also known from previous research that students in problem-based schools show more deep learning behaviour and are more actively involved in the learning process compared to students in traditional schools.[31-33] With increasing years of study the percentages of students reporting surface-active and surface-passive learning behaviour fell whereas the percentage of students reporting deep-active learning behaviour rose. These findings are comparable with earlier research.[31] Whether this is due to an effect of PBL, the PT or personal development of the students cannot be determined. We also found a trend towards an increase with years of study in the contribution of desired learning behaviour to the explained variance in PT scores. In the fourth year, 30% of PT score variance could be explained by desired problem based learning behaviour. Why this should be so is difficult to explain, but it might reflect a positive effect of deep learning behaviour. With increasing years of study the cumulative effects of learning behaviours could result in a greater divergence of PT scores between students who use different learning strategies.

Prediction of academic achievement from learning behaviour has been a popular topic of research. Studies which included many variables concerning learning behaviour explained only a limited amount of variance in achievement.[34,35] Considering the number of variables used in this study and the contrasting findings for UT and OSCE, the incremental explanation of the variance in PT scores by problem based learning behaviour is a striking finding.

Test directed learning and test preparation
Test directed preparation only seems to pay off for the OSCE. Learning of motor skills requires practice. Because the OSCE tests "hands-on" performance, it is not surprising that preparation should positively affect test scores. Earlier studies confirmed the positive relationship between practice and OSCE scores.[36] This is definitely not true for the number of hours studied for the UT. Although about 53% of the students were found to spend an average of 20 hours on test-directed preparation for the UT, this appeared to offer no guarantee of good test results. On the contrary, specific preparation for the UT was

negatively correlated with all test scores including the UT score. This finding is remarkable considering that the UT tests the content of a six-week course. Thus the content of the UT is easy to predict and test directed learning might be expected to yield better test results. However, a greater percentage of study time spent on specific test preparation was correlated with fewer hours of general study and less desirable learning behaviour. This may explain the inverse relationship between test-directed study and (UT) scores. Students reported less test-directed studying in preparation for the PT compared to the UT. Also, fewer students prepare for the PT (35%). This finding confirms the results of earlier research in which 32% of students were found to prepare for the PT.[37,38] This percentage is higher than that reported at McMaster[38,39] and the percentage has risen between 1980 and 1995.[38,39] The number of hours students claim to spend on preparation is limited, only three hours per PT. This is insignificant compared to the hours spent on UT and OSCE preparation. Test-directed preparation for the PT is not associated with higher PT scores. Why and how students prepare for the PT has to be investigated in future research.

Conclusion

Students who reported desirable learning behaviour scored higher on the PT compared to students with less desirable learning behaviour. Low scoring students reported more passive and surface learning behaviour. Learning behaviour explained 10% of the variance in PT scores. The number of hours students claimed to spend preparing for the PT was limited and did not appear to affect test results. On the whole, the results of this study suggest that progress testing achieves its purpose and thus provide evidence for the consequential validity of the test.

References

1 Marton F, Säljö R. On qualitative differences in learning: II - Outcome as a function of the learner's conception of the task. *British Journal of Educational Psychology* 1976; **46**: 115-27.

2 Hilliard RI. How do medical students learn: Medical students learning styles and factors that affect these learning styles. *Teaching and Learning in Medicine* 1995; **7**: 201-10.

3 Marton F, Säljö R. On qualitative differences in learning: I - Outcome and process. *British Journal of Educational Psychology* 1976; **46**: 4-11.

4 Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Medical Education* 1983; **17**: 165-71.

5 Frederiksen N. The real test bias. Influences of testing on teaching and learning. *American Psychologist* 1984; **39**: 193-202.

6 Cohen-Schotanus J. Student assessment and examination rules. *Medical Teacher* 1999; **21**: 318-21.

7 Van der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education* 1996; **1**: 41-67.

8 Miller GE. Continuous assessment. *Medical Education* 1976; **10**: 81-6.

9 Newble DI, Entwistle NJ. Learning styles and approaches: Implications for medical education. *Medical Education* 1986; **20**: 162-75.

10 Verwijnen M, Imbos T, Snellen H, Stalenhoef B, Pollemans M, Van Luyk S, Sprooten M, Van Leeuwen Y, Van der Vleuten C. The evaluation system at the medical school of Maastricht. *Assessment and Evaluation in Higher Education* 1982; **7**: 225-44.

11 Arnold L, Willoughby TL. The quarterly profile examination. *Academic Medicine* 1990; **65**: 515-6.

12 Van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher* 1996; **18**: 103-9.

13 Blake JM, Norman GR, Kinsey E, Smith M. Report card from McMaster: Student evaluation at a problem-based medical school. *Lancet* 1995; **345**: 899-902.

14 Ebel RL, Frisbie DA. *Essentials of educational measurement.* 5th ed. New Jersey: Englewood Cliffs, 1991.

15 Moss PA. Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research* 1992; **62**: 229-58.

16 Messick S. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 1994; **23**: 13-23.

17 Sambell K, McDowell L, Brown S. "But is it fair?": An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation* 1997; **23**: 349-71.

18 Van der Vleuten CPM, Scherpbier AJJA, Wijnen WHFW, Snellen HAM. Flexibility in learning: A case report on problem-based learning. *International Higher Education* 1996; **2**: 17-24.

19 Van der Vleuten C, Verwijnen M. A system for student assessment. In: Van der Vleuten C, Wijnen W, editors. *Problem-based learning: Perspectives from the maastricht experience.* Amsterdam: Thesis; 1990: 27-49.

20 Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education* 1979; **13**: 41-54.

21 Biggs J, Kember D, Leung DY. The revised two-factor study process questionnaire: R-SPQ-2F. *British Journal of Educational Psychology* 2001; **71**: 133-49.

22 De Grave WS, Boshuizen HPA, Schmidt HG. Problem based learning: Cognitive and metacognitive processes during problem analysis. *Instructional Science* 1996; **24**: 321-41.

23 Hol TPM. Persoon - situatie interacties: Operationalisering, gedragsvoorspelling en modelvergelijking [Person - situation interactions: Operationalization, prediction of behaviour and comparison of models] [PhD dissertation Katholieke Universiteit Brabant]. Tilburg, 1994.

24 Van Til CT. Voortgang in voortgangstoetsing. Studies naar de aansluiting van de voortgangstoets op probleemgestuurd onderwijs [Progress in progress testing. Studies on the suitability of progress testing within a problem-based educational context] [PhD dissertation Maastricht University]. Wageningen: Ponsen & Looijen, 1998.

25 Van Til CT, Van der Vleuten CPM, Van Berkel HJM. Problem-based learning behavior: The impact of differences in problem-based learning style and activity on students' achievement. *Annual Meeting of the American Educational Research association.* Chicago: 1997. ERIC No. TM026783 (ED409333).

26 Schmidt HG. Problem-based learning: Rationale and description. *Medical Education* 1983; **17**: 11-6.

27 Schmidt HG. Foundations of problem-based learning: Some explanatory notes. *Medical Education* 1993; **27**: 422-32.

28 Norman GR, Streiner DL. *Biostatistics: The bare essentials*. 2nd ed. Hamilton: B.C. Decker Inc., 2000.
29 Van Luijk SJ, Melief RM. Toetsresultaten en studiestijlen [Test results and learning styles]. In: Van der Vleuten CPM, Scherpbier AJJA, Pollemans MC, editors. *Gezond onderwijs 1*. Houten / Diegem: Bohn Stafleu Van Loghum; 1992: 111-6.
30 Van Berkel HJM, Nuy HJP, Geerlings T. The influence of progress test and block tests on study behaviour. *Instructional Science* 1995; **22**: 317-33.
31 Newble DI, Clarke RM. The approaches to learning of students in a traditional and in an innovative problem-based medical school. *Medical Education* 1986; **20**: 267-73.
32 Coles CR. Differences between conventional and problem-based curricula in their students' approaches to studying. *Medical Education* 1985; **19**: 308-9.
33 Albanese MA, Mitchell S. Problem-based learning: A review of literature on its outcomes and implementation issues. *Academic Medicine* 1993; **68**: 52-81.
34 Van Overwalle F. Success and failure of freshmen at university: A search for determinants. *Higher Education* 1989; **18**: 287-308.
35 Meerum Terwogt-Kouwenhoven K. Niet gewogen, toch te licht bevonden [Not assessed, but too feeble; analyses of output problems in Dutch higher education] [PhD dissertation University of Amsterdam]. Amsterdam, 1990.
36 Van Venrooij EJ, Verhoeven BH, Van Zandvoort HPHC, Bartholomeus PMTA, Scherpbier AJJA. Het docent-onafhankelijk oefenen van vaardigheden. Een exploratieve studie [Unsupervised skills training. An exploratory study]. *Bulletin Medisch Onderwijs* 1998; **17**: 31-6.
37 Brümmer I, Linden ML, Hoogenboom RJI, Scherpbier AJJA, Verwijnen GM. Het oordeel over de Maastrichtse voortgangstoets. Het consumentenoordeel in 1995 [Assessment of the Maastricht progress test. Consumer opinion 1995]. *Bulletin Medisch Onderwijs* 1995; **14**: 174-85.
38 Linden ML, Brümmer I, Hoogenboom RJI, Scherpbier AJJA, Verwijnen GM. De Maastrichtse voortgangstoets. Een vergelijking van drie peilingen van het consumentenoordeel [The Maastricht progress test. A comparison of three samplings of consumer judgement]. *Bulletin Medisch Onderwijs* 1995; **14**: 186-91.
39 Blake JM, Norman GR, Keane DR, Mueller CB, Cunnington J, Didyk N. Introducing progress testing in McMaster university's problem-based medical curriculum: Psychometric properties and effect on learning. *Academic Medicine* 1996; **71**: 1002-7.

Chapter 10

An analysis of progress test results of PBL and non-PBL students

Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Holdrinet RSG, Oeseburg B, Bulte JA, Van der Vleuten CPM
Published in *Medical Teacher* 1998; **20**: 310-6

Introduction

Since the introduction of Problem-Based Learning (PBL), many studies have been reported comparing the effectiveness of PBL curricula with more traditional educational programs.[1-3] One aspect of interest is the level of factual medical knowledge that graduates of a problem-based learning curriculum have in comparison to graduates from more traditional schools. Most research in this respect uses student achievement results on licensure examinations such as the USMLE (formerly NBME). One part of these exams covers basic sciences and most students take this part at the end of the second year. Another part covers the clinical sciences and is most often taken during the fourth year.[4-6] In so far as differences occurred, PBL students had lower test scores on the basic sciences part and better test scores on the clinical sciences part.[1,2] Several other examinations required for licensure and graduation, including oral exams and clinical rotation ratings, have also been used for comparisons. The differences found between PBL and non-PBL were not very consistent. The selection of students and implementation differences in PBL probably account for these inconsistent findings.[7] The fact that some studies mention differences between PBL and non-PBL students in knowledge of basic sciences at the end of the second year does not have to mean that PBL students eventually have less knowledge of basic sciences at the end of the curriculum.[8] It is plausible that the gathering of specific medical knowledge does not occur at the same moment in time in a non-PBL curriculum as in a PBL curriculum.

In this study the medical factual knowledge of PBL and non-PBL students is compared at all levels of training. The major differences between the two medical schools involved are found in the first four years. The PBL school offers a problem-based, student-centred and integrated curriculum that is organized in themes. The program consists of interdisciplinary units of usually six weeks.[9] More attention is paid to basic and social sciences during the first two years than in years 3 and 4. The clinical sciences dominate in years 3 and 4. At the time this study was conducted, the non-PBL school had a conventional discipline-oriented educational method marked by instructor-provided learning objectives and assignments, large group lectures and structured laboratory experiences. The teaching of the principles of basic sciences (mainly years 1 and 2) preceded the instruction in clinical sciences (mainly years 3 and 4). In the last two years all Dutch medical students spend their time on clinical rotations and no major program differences exist between the medical schools.

At the time of the study a curriculum change was in preparation in the non-PBL school. As part of the preparation, the PBL school's Progress Test was administered to students of all six classes of the non-PBL school. This was done to get experience with administration of this kind of test and to provide a base-line for further research guiding the curriculum change. Ideally a random sample of students had to participate, but this could, for practical reasons not be realised. Therefore volunteers were asked to make the PT's. Volunteers are mostly the better students, but this methodological shortcoming had to be taken for granted.

We were interested whether there would be differences in students' results on the progress test with regard to 1) total test score (overall medical knowledge); 2) subscores in basic sciences,

clinical sciences and social sciences; 3) subscores on disciplines (anatomy, surgery, ENT, et cetera) in all curriculum years; and 4) scores on individual test-items.

Methods

Instrument

The results on the progress test were used in this comparison.[10] The PT is best characterized as a comprehensive final examination in medicine. Each PT samples the complete domain of factual knowledge that is considered pertinent for a medical graduate to master and consists of approximately 250 true/false questions with a question mark ("I do not know") option. These questions are stratified in categories based on the International Classification of Diseases (ICD). The content of the questions may or may not be covered by the educational programme. Therefore the PT can be called a "programme independent" test and can be administered at every medical school regardless of its specific educational programme. The PT is given four times a year to all students in the curriculum regardless of their class. An equivalent test is produced for each occasion on the basis of a blueprint. In this way it is possible to monitor progress or growth towards the overall objectives.[8,11,12] Two progress tests were used in this study: the PT of December 1994 consisting of 242 items; the PT of March 1995 comprising of 226 items.

Subjects

Students of all six classes of two medical schools in the Netherlands, one PBL and one non-PBL, were used (the medical training program takes six years in the Netherlands). The freshmen of the eight medical schools are comparable with respect to age, gender and knowledge base. They enter medical school directly after secondary education. The secondary school system in The Netherlands is quite homogeneous and students only graduate after passing a national secondary school examination. Medical school admission is centralized. The selection procedure is partly based on grade point average and partly on a lottery, therefore called a "weighted" lottery system.[13] After this selection, students are assigned to one of the eight medical schools in The Netherlands. No academic differences between entering medical students across schools are to be expected through this system.

Table 1 shows the number of participating students from both schools per class as well as the percentage of the participating non-PBL students in relation to the total number of students in their year.

Table 1.
Number of participating students on two administrations of the Progress Tests (percentages of non-PBL classes in parentheses).

	December 1994			March 1995		
Year	PBL	non-PBL	(% of class)	PBL	non-PBL	(% of class)
1	190	124	(68%)	218	54	(30%)
2	146	104	(64%)	151	60	(37%)
3	135	87	(49%)	133	51	(29%)
4	188	151	(58%)	169	53	(20%)
5	144	140	(68%)	155	69	(33%)
6	135	122	(59%)	140	74	(36%)

Procedure
Two PTs were administered. At the PBL school the PT is part of the regular assessment program and the test results had consequences for study progress. At the non-PBL school the PT was not part of the assessment program. All participating students were volunteers and their test scores had no consequences for advancement decisions. We had no information on the representativeness of the volunteer group. The two administrations used in this study took place in December 1994 and in March 1995. On both occasions it concerned fully synchronized administrations at the same moment in time only at different places.

Statistical analysis
To discourage guessing, formula scoring was used in which the incorrect scores are subtracted from the correct scores (the "I don't know" option receives no marks). For each student the scoring was expressed as a percentage of the total number of questions. For both PTs the means and standard deviations were calculated per class and per faculty. Total scores as well as subscores on basic sciences (anatomy, biochemistry, pharmacology, physiology, genetics & cell biology, immunology, microbiology, and pathology), clinical sciences (surgery, cardiology, dermatology, gynecology & obstetrics, family medicine, internal medicine, pediatrics, ENT, neurology, orthopedics, ophthalmology, pulmonology and rehabilitation medicine) and social sciences (health care economy, epidemiology, health care law, ethics & philosophy, medical psychology, medical sociology, and social psychiatry) were compared. To test the significance of the differences between the mean total scores and subscores per year group of both schools, a one-way ANOVA was used. To correct for the number of comparisons, a Bonferroni correction was used and a $p = 0.01$ was considered to be significant. The correct minus incorrect scores on disciplines and items were compared using visual inspection of plots and correlation coefficients (Pearson product-moment correlation coefficient).

Results

In table 2 the correct minus incorrect scores of both schools are presented per year on the complete test as well as on the three clusters of disciplines (basic, clinical and social sciences). The statistically significant differences in scores found in December 1994 do not correspond with those found in March 1995. Only the differences that are found on the basic sciences scores in the first and sixth years are significantly different in both December and March. Remarkably, however, this difference does not exist for year five. In March 1995 the significant differences mainly concern the basic sciences in favours of the non-PBL group, whereas in December 1994 the significant differences are found in the social sciences domain in favour of the PBL-group.

At the discipline level of comparison, each discipline may be presented by only a few test items, causing a considerable spread of scores. For this reason statistical significance testing has not been used and a visual inspection of the profile was preferred. In figures 1a to 1d the mean test score on each discipline is depicted for the first and sixth year separately. The parallelism of the profile of the curves is striking. There is quite some variation across disciplines, but this is (almost completely) identical for both schools. Interestingly, the profiles are not the same across both test administrations.

The most detailed level of comparison is a comparison of the "itemscores". These scores represent the percentage of students who answered a question correctly minus the percentage of students that answered the same question incorrectly. These "item scores" (242 test-items in December and 226 in March) of both faculties are plotted in scatter diagrams for years one, three and six (figures 2 to 4). Data points on the diagonal represent perfect resemblance between the scores of the PBL and non-PBL students. The more distant a data point is from the diagonal, the less resemblance is present between the two schools on that specific item, suggesting that the PBL and non-PBL students have different (curriculum-specific) knowledge.

Table 2.
Students' mean correct minus incorrect scores and Standard Deviations (SD) on total test and clusters of disciplines.

	December 1994					March 1995				
		PBL		non-PBL			PBL		non-PBL	
Year	No. of items	Mean	SD	Mean	SD	No. of items	Mean	SD	Mean	SD
Total Test	242					226				
1		9.61	4.06	10.70	5.37		11.42	4.29	12.17	6.85
2		20.68	5.94	19.27	6.68		20.64	6.43	22.26	8.27
3		26.18	6.68	23.50	9.32		26.51	6.88	30.37†	9.14
4		33.27	7.98	31.54	8.95		35.14	7.64	38.03	9.28
5		38.46	8.53	38.80	9.16		38.41	8.59	41.52	9.37
6		41.03	8.01	44.49†	8.99		42.12	6.71	44.85	11.07
Basic Sciences	92					93				
1		8.43	4.77	10.68†	6.78		11.26	5.66	14.97†	8.20
2		19.09	8.27	16.66	8.59		18.78	8.19	25.48†	10.74
3		21.08	8.56	19.63	10.00		24.46	8.88	31.58†	11.06
4		25.86	9.31	25.66	10.50		32.46	8.96	37.74†	10.02
5		30.13	10.97	31.16	11.70		36.25	10.39	38.41	10.30
6		32.33	10.21	37.07†	11.00		36.49	8.24	41.08†	12.09
Clinical Sciences	103					91				
1		8.24	4.74	9.23	6.65		4.77	4.57	3.74	7.90
2		19.26	6.94	20.46	8.13		14.41	7.46	13.33	8.53
3		26.39	8.62	26.09	10.60		23.10	8.69	25.02	10.99
4		37.88	9.80	36.15	10.80		34.38	10.33	35.41	11.92
5		44.56	9.71	45.30	10.40		38.72	11.06	41.77	13.35
6		48.71	10.04	52.51†	11.10		46.44	9.12	48.60	12.85
Social Sciences	47					42				
1		14.93	9.74	13.98	10.30		26.18	11.90	24.21	14.33
2		26.90	11.20	21.77†	10.40		38.25	12.40	34.44	12.78
3		35.70	12.71	25.39†	13.80		38.42	12.67	39.26	12.57
4		37.65	12.86	32.93†	13.90		42.72	12.25	44.34	14.42
5		41.43	13.04	39.53	12.10		42.53	13.40	47.83†	13.19
6		41.26	11.36	41.42	12.37		45.24	11.15	45.05	16.74

†Statistical significance at the level $p < 0.01$ (one-way ANOVA).

Figure 1.
Disciplinary knowledge profiles (mean correct minus incorrect score on each discipline) of PBL and non-PBL students. A) First year students, Progress Test of December 1994; B) first year students, Progress Test of March 1995; C) sixth year students, Progress Test of December 1994; D) sixth year students, Progress Test of March 1995.

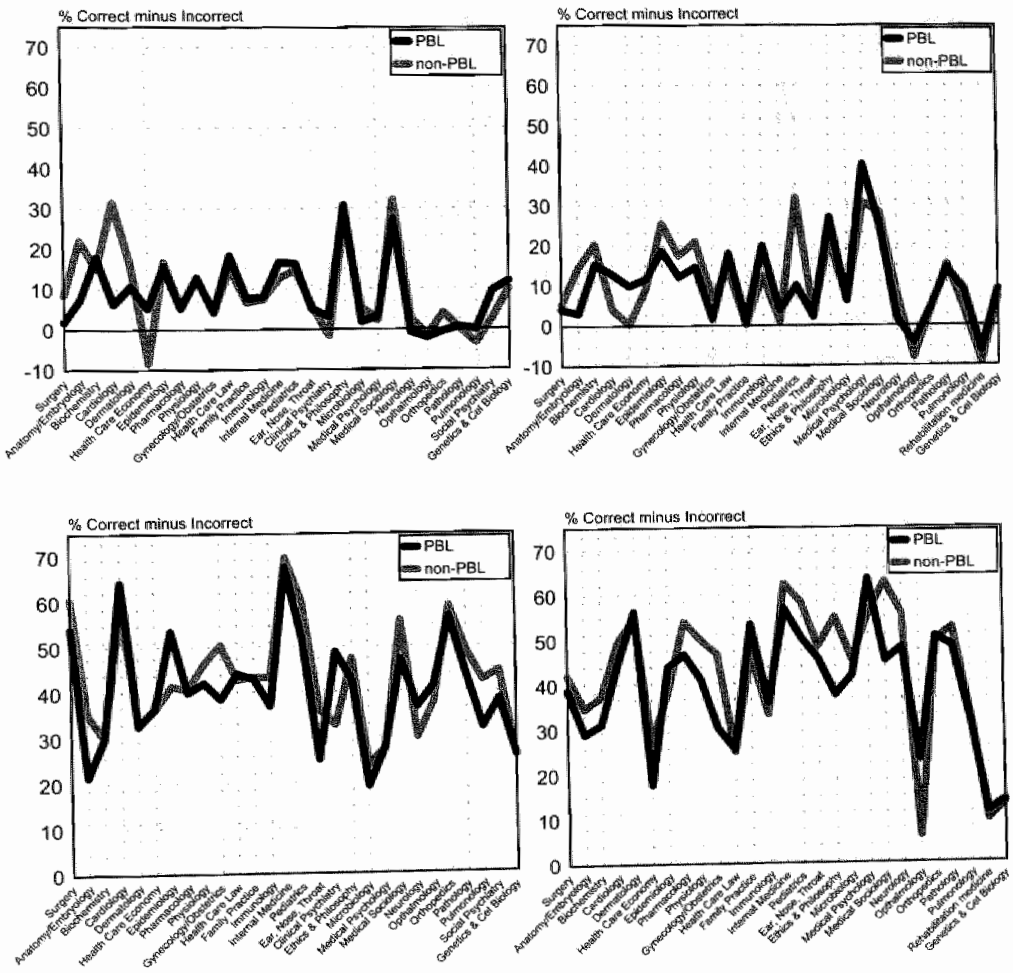


Figure 2.
Comparison of correct minus incorrect "itemscores" of first year PBL and non-PBL students. A) All 242 items of Progress Test of December 1994; B) All 226 items of Progress Test of March 1995.

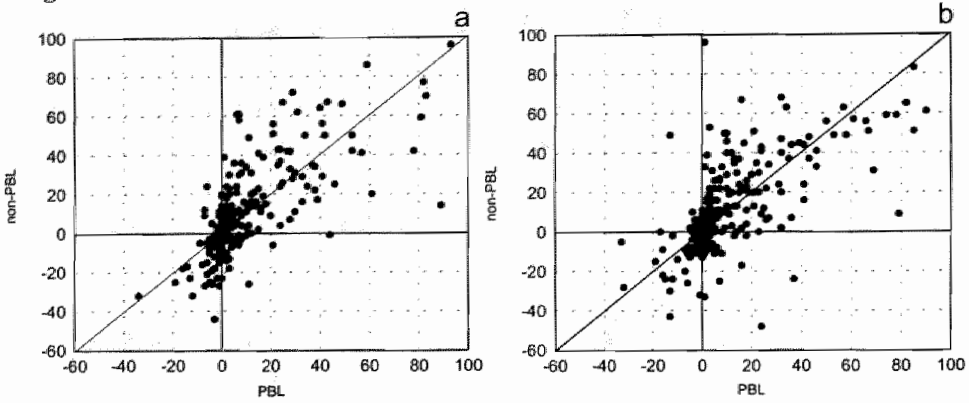
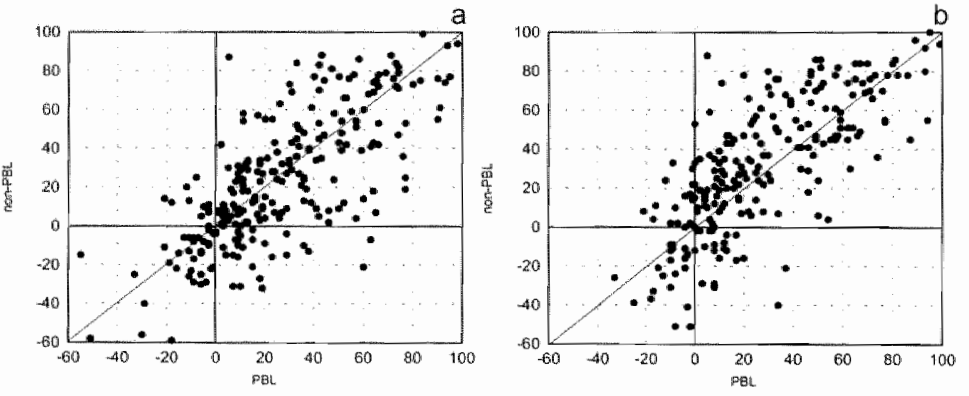


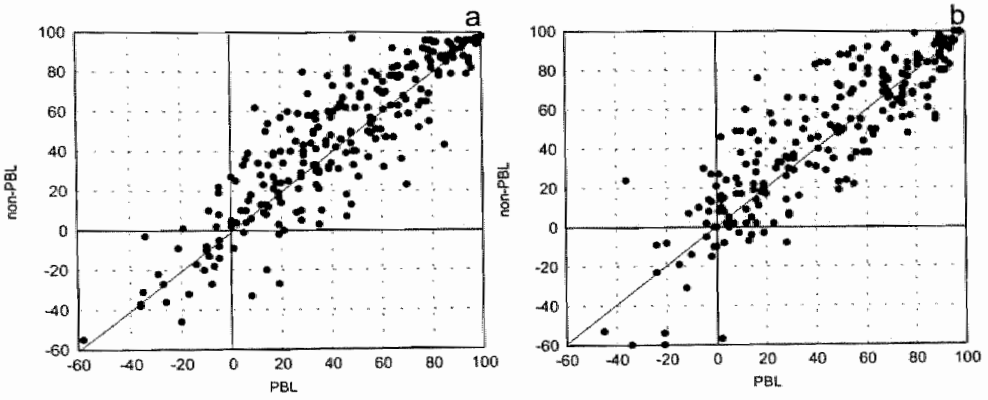
Figure 2 contains the data of year one. A cluster of data-points is situated near the intersection of the two axis of the graph, meaning that many items remain unanswered or that many students in both faculties do not have the knowledge to answer those items correctly. Some items are situated near the diagonal and further away from the intersection of the axis. Students of both faculties have the same level of knowledge on these items. In addition, some items are answered correctly by either more non-PBL or more PBL students, suggesting the existence of curriculum-specific knowledge on these items.

Figure 3.
Comparison of correct minus incorrect "itemscores" of third year PBL and non-PBL students. A) All 242 items of Progress Test of December 1994; B) All 226 items of Progress Test of March 1995.



In the third year (figure 3), fewer data points are situated near the intersection, meaning that students, as can be expected, answer more questions. However, quite a lot of items are answered correctly by either more non-PBL or more PBL students. This suggests curriculum-specific knowledge on those questions. The number of items that have a negative correct minus incorrect score is substantial, meaning the students with faulty knowledge outnumber those with correct knowledge.

Figure 4.
Comparison of correct minus incorrect "itemscores" of sixth year PBL and non-PBL students. A) All 242 items of Progress Test of December 1994; B) All 226 items of Progress Test of March 1995.



In the final year of medicine (figure 4) a small cluster of items is situated in the upper right corner of the diagram; almost all students of both faculties possess the knowledge to answer these questions correctly. The scatter of data-points is less wide and most items are situated near the diagonal, meaning there is a considerable resemblance between the scores of the PBL and non-PBL students. The number of items that have a negative score is diminished. The omitted years (2, 4 and 5) show perfect intermediate positions. The correlations between the item scores of PBL and non-PBL students are depicted in table 3. In the course of years, the correlation rises, suggesting the differences between PBL and non-PBL students diminish.

Table 3.
Pearson correlation coefficients of item scores of PBL and non-PBL students on two administrations of the Progress Test.

Year	December 1994	March 1995
1	0.72	0.66
2	0.69	0.69
3	0.73	0.76
4	0.83	0.78
5	0.87	0.85
6	0.88	0.86

Discussion

The results obtained in this study strongly agree with those of the previous comparisons between medical schools.[4,7,8,11,12,14] In this study the similarities in level of cognitive performance again prevail over the differences. As was expected, no systematic differences are found on total test scores. After splitting the test in the three categories (basic, clinical and social sciences), a few, but non-systematic differences were found. To some extent the non-PBL students scored better on the basic sciences while the PBL students had better scores on the social sciences, but this was not consistent for both test administrations. These findings are quite surprising given the apparent and substantial differences between the two curricula involved.[15,16] It is also quite amazing that the PBL students in the first and second years failed to outperform the non-PBL students on clinical sciences given the integrated nature of their educational program.

The parallelism in results is even found at a more detailed level of comparison. The mean scores on the disciplines show remarkable resemblance. Apparently questions from a discipline are more difficult or easier irrespective of the school's training method. The level of difficulty of a cluster of questions within a certain discipline or category appears to be rather a generic phenomenon, varies from test to test, and is independent of the curriculum that one follows.

Only at the level of individual questions differences between PBL and non-PBL could be demonstrated. In the first year of medical school many questions can be found that are answered better by PBL or non-PBL students. However this effect diminishes systematically with increasing seniority in the curriculum. At the end of the sixth year these differences at item level too yielded nearly total parallelism. At the end of their curricula, PBL and non-PBL students answer the same questions correctly and incorrectly. The students master the same knowledge; it is only the moment in time that they learn it that differs.

Methodological considerations

A few methodological considerations can be made with regard to this study. In the non-PBL medical school all participating students were volunteers. The test results had no consequences for them in contrast with the PBL students, for whom the tests were obligatory and part of their exams. Both aspects are essential methodological shortcomings of this study and make a valid interpretation of this comparison more difficult.[17] Volunteers in most cases are the better students[18] and tests that have no consequences for the participators lead to different answer strategies, particularly more guessing. Indeed, we found that the volunteers (non-PBL students) did use significant fewer question marks (data not shown). To correct for these different answer strategies, formula scoring was used in this comparison.

Perhaps the questions that are used in the PT are not sensitive enough to detect differences between curricula. They could be too rough, not tailored enough, asking only those knowledge aspects that everyone will learn anyway by spending six years in a medical school. The real differences could be in topics that are not included in the PT and therefore are still to be discovered. Research, however, shows that the PBL students experience the PT as a relevant test that reflects topics covered by their curriculum.[19] This, plus earlier comparisons, suggests the PT is a valid and valuable instrument to compare medical curricula, even internationally.[20]

Despite these methodological shortcomings, this study indicates that the effects of PBL and non-PBL instructional methods on medical factual knowledge output are even more similar than we had previously thought.

References

1 Vernon DTA, Blake RL. Does problem-based learning work? A meta-analysis of evaluative research. Academic Medicine 1993; 68: 550-63.

2 Albanese MA, Mitchell S. Problem-based learning: A review of literature on its outcomes and implementation issues. Academic Medicine 1993; 68: 52-81.

3 Berkson L. Problem-based learning: Have the expectations been met? Academic Medicine 1993; 68: S79-88.

4 Schmidt HG, Dauphinee WD, Patel VL. Comparing the effects of problem-based and conventional curricula in an international sample. Journal of Medical Education 1987; 62: 305-15.

5 Swanson DB, Case SM, Melnick DE, Volle RL. Impact of the USMLE step 1 on teaching and learning of the basic biomedical sciences. Academic Medicine 1992; 67: 553-6.

6 National Board of Medical Examiners. 1995 performance on USMLE examinations, <http://nbme.org/step95.htm>, accessed 25 March 1997.

7 Woodward CA. Problem-based learning in medical education: Developing a research agenda. Advances in Health Sciences Education 1996; 1: 83-94.

8 Van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. Medical Teacher 1996; 18: 103-9.

9 Van der Vleuten CPM, Scherpbier AJJA, Wijnen WHFW, Snellen HAM. Flexibility in learning: A case report on problem-based learning. International Higher Education 1996; 2: 17-24.

10 Verwijnen M, Imbos T, Snellen H, Stalenhoef B, Pollemans M, Van Luyk S, Sprooten M, Van Leeuwen Y, Van der Vleuten C. The evaluation system at the medical school of Maastricht. Assessment and Evaluation in Higher Education 1982; 7: 225-44.
11 Verwijnen M, Van der Vleuten C, Imbos T. A comparison of an innovative medical school with traditional schools: An analysis in the cognitive domain. In: Nooman ZM, Schmidt HG, Ezzat ES, editors. Innovation in medical education: An evaluation of its present status. Vol. 13. New-York: Springer Publishing Company; 1990: 40-9.
12 Van Hessen PAW, Verwijnen GM. Does problem-based learning provide other knowledge? In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, editors. Teaching and Assessing Clinical Competence. Groningen: 1990. 446-51.
13 Wijnen WHFW. The Netherlands. In: Burn BB, editor. Admission to medical education in ten countries. New York: International Council for Educational Development; 1978: 105-14.
14 Bender W, Cohen-Schotanus J, Imbos T, Versfelt WA, Verwijnen GM. Medische kennis bij studenten uit verschillende medische faculteiten: Van hetzelfde laken een pak? [Medical knowledge in students from various medical schools: Being served with the same sauce?]. Nederlands Tijdschrift voor Geneeskunde 1984; 128: 917-21.
15 Snellen-Balendong HAM, Wijnen WHFW, Langevoort HLL. Diversiteit van medische curricula in nederland. [Diversity of medical curricula in The Netherlands]. Nederlands Tijdschrift voor Geneeskunde 1994; 138: 1136-42.
16 Faculteit der Medische Wetenschappen. Studiegids geneeskunde 1994/1995 [Study guide medicine 1994/1995]. Nijmegen, The Netherlands: Universiteitsdrukkerij, 1994.
17 Schmidt HG. Innovative and conventional curricula compared: What can be said about their effects? In: Nooman ZM, Schmidt HG, Ezzat ES, editors. Innovation in medical education. An evaluation of its present status. New York: Springer Publishing Company; 1990: 1-7.
18 Ten Cate TJ. De invloed van anonimiteit op de resultaten van tijdschrijfonderzoek [The influence of anonymity on the results of time registration research]. Tijdschrift voor Onderwijsresearch 1985; 10: 263-73.
19 Linden ML, Brümmer I, Hoogenboom RJI, Scherpbier AJJA, Verwijnen GM. De Maastrichtse voortgangstoets. Een vergelijking van drie peilingen van het consumentenoordeel [The Maastricht progress test. A comparison of three samplings of consumer judgement]. Bulletin Medisch Onderwijs 1995; 14: 186-91.
20 Albano MG, Cavallo F, Hoogenboom R, Magni F, Majoor G, Manenti F, Schuwirth L, Stiegler I, Van der Vleuten C. An international comparison of knowledge levels of medical students: The Maastricht progress test. Medical Education 1996; 30: 239-45.

Chapter 11

The versatility of progress testing assessed in an international context A start for benchmarking global standardization?

Verhoeven BH, Snellen-Balendong HAM, Hay IT, Boon JM, Van der Linde MJ, Blitz JJ, Hoogenboom RJI, Verwijnen GM, Wijnen WHFW, Scherpbier AJJA, Van der Vleuten CPM
Under editorial review

Introduction

Progress testing is a longitudinal form of testing knowledge in a comprehensive way.[1] Periodically, i.e. 4 times a year, a newly written but parallel test is administered to all the students or residents from a particular programme at the same time. First year candidates will not be able to answer many questions, second year candidates somewhat more and so on until a final level is reached. The test consists of a large number of multiple choice or true/false questions covering all disciplines within the curriculum. The progress test (PT) was originally developed to assess knowledge in a problem based learning context to avoid test directed studying and to reinforce self directed learning.[1] All individual study activities will be rewarded by the test. Extensive experience and empirical research provided evidence that the progress test works in that way.[1-6] Progress testing is a valuable assessment tool in any context emphasizing self directed learning. PTs are currently used in a number of undergraduate medical schools and some postgraduate training programmes.[4,7-12]

Another advantage is that the test procedure is not course specific but comprehensive and intended to reflect the final objectives of a curriculum. Progress tests in a medical undergraduate domain, for example, are stratified by agreed standards and tables of specification.[13,14] These specifications are not confined to one specific medical curriculum. Therefore the test can easily be applied in other medical schools.[15] This allows not only to compensate a disadvantage of progress testing, but delivers yet an advantage. This potential expanded application of the PT would underline its value as a comparative assessment instrument and help to overcome the resource intensity drawback. The disadvantage of progress testing is the resources that are required to develop and maintain an item bank with sufficient new and high quality questions.[13] So why not share and collaboratively develop and maintain such a bank across institutions? Three medical schools in The Netherlands collaboratively develop progress test questions.[16] A fourth school participates by buying the test. The same test is administered in all four schools at the same time. Thereby the progress testing procedure allows for quality control. It is possible to compare subjects and schools and evaluate growth of knowledge.[13,17-19] This versatile characteristic of the progress test might be a good step on the way to more global standardization of high quality tests.[20-22] The items in the PT are the materialized objectives of a (core) curriculum and the performance of the group of candidates are the realized objectives. Cross institutional sharing and collaboration, and quality control are easily achieved with this instrument.

PTs can be administered across different schools within a particular country. Some time ago it was demonstrated that PT's can be used to compare results of candidates of different countries.[23] A PT was administered in a German, a Dutch and four Italian medical schools. The study demonstrated the feasibility of an international progress testing approach, but the performance levels were sometimes rather variable and difficult to interpret. However, only a limited attempt was made to correct for cultural bias and specific problems resulting from the translation of the original Dutch test. In the current study we wanted to make a

comparison holding better control on these sources of bias. The results of this study will provide information how far PTs and PT results are exchangeable between different countries and cultures. This way the potential possibilities of the PT as one of the instruments for world-wide quality control of medical education are explored.

Methods

Participating institutions

This comparison involved medical students from the University of Pretoria in South Africa and the University of Maastricht in The Netherlands. Until the intake of January 1997, the medical curriculum in Pretoria was a traditional, discipline based, lecture-based six year programme. In the last year clerkships were scheduled in the four major clinical specialities (internal medicine, surgery, paediatrics, obstetrics and gynaecology), as well as anaesthesiology. In 1997 a new integrated community-based, problem-oriented curriculum was introduced aimed at improving the relevance of the curriculum to the practice of medicine. Horizontal and vertical integration of content is pursued in which the students are exposed to clinical teaching from a much earlier point in the curriculum.[24] The final clerkships run for a period of 18 months in the previously mentioned disciplines as well as psychiatry and family medicine. This training aims to prepare students for the next two years of their careers which will be public sector, largely hospital- and speciality-based practice of a 1 year internship and 1 year of compulsory community service. At the time of the study, students were registered in both the old and new curricula. The medical school of the University of Maastricht has a six-year undergraduate programme leading to a medical degree which allows application to further postgraduate training in the specialities. The PBL programme is horizontally and vertically integrated.[25,26] Self-directed learning is considered important. The number of educational contact hours scheduled per week is about 10 to 12 hours, consisting predominantly of tutorial group sessions and Skillslab activities. The last two years consist of clerkships.

Instrument

Three Maastricht progress tests (March 1997, May 1997 and September 1997) were translated into English and Afrikaans by one of the authors. Afrikaans stems from Dutch originally. Therefore Dutch could be easily understood and accurately translated by the Pretoria doctor The translation was reviewed by a second person and the Evaluation Committee. Students in Pretoria could choose to read the English or the Afrikaans version. Both versions were in one booklet. The original tests were based on the regular blueprint accepted by the collaborating Dutch medical schools.

Subjects

The tests were administered in all years of training with a few exceptions. First year students in Pretoria did not participate in the March and September test. Freshmen are located on a campus 12 km away from the test administration site and only wrote the May test. Year 6 students were not required to participate in the May test because they were

working as clerks and therefore off campus scattered at different teaching sites. The first to fifth year students still attend lectures and were all on the campus. Table 1 shows the number of participating students from both schools per class as well as the percentage of the participating Pretoria students in relation to the total number of students in their year.

Table 1.
Number of participating students on three administrations of the progress test (percentages of the Pretoria classes in parentheses).

	March 1997 June 1999			May 1997 September 1998			September 1997 July 1998		
Year	M	P	(% class)	M	P	(% class)	M	P	(% class)
1	202	0	(0)	192	204	(100)	179	0	(0)
2	192	178	(86)	189	207	(100)	191	206	(100)
3	220	220	(94)	215	198	(98)	199	222	(100)
4	153	221	(100)	117	197	(86)	218	195	(85)
5	134	202	(100)	136	203	(98)	153	196	(95)
6	108	185	(86)	69	0	(0)	134	209	(100)

Procedure
The translated tests were administered to the medical students in Pretoria in 1998 and 1999. In Pretoria the academic year starts in January, while in Maastricht the year opens in September. Table 2 shows the months in which the three tests were administered in Pretoria. In the same table the moment of test administration is related to the start of the academic year.

After test administration in Pretoria, one Dutch medical doctor, familiar with the Pretoria setting through several visits to South Africa, reviewed the translated items on potential cultural bias. The reviewer was instructed to remove an item even with the least suspicion of a translation error or cultural bias, i.e. items that do not have applicability outside the Dutch society. After the culled questions were removed, the test results were further analysed.

Table 2.
The months in which the PTs were administrated in the two participating faculties. On the right half of the table, the number of months after the start of the academic year is depicted.

Test	Moment of administration		Months after start of academic year	
	Maastricht	Pretoria	Maastricht	Pretoria
1	March 1997	June 1999	7	6
2	May 1997	September 1998	9	9
3	September 1997	July 1998	1	7

Analysis
A progress test item score sheet allows for not answering the question by completing an "I don't know" option. This is encouraged by the scoring system. For each correct answer one point is given, for each incorrect answer a point is subtracted. Question mark ("I don't know") answers receive no credits. The final scoring is the total number of correct minus the number of incorrect (so-called formula scoring). For each student the correct minus incorrect score was calculated and expressed on a percentage scale. This was done for the total score and for the scores on the three clusters of disciplines: basic sciences (anatomy, biochemistry, pharmacology, physiology, genetics & cell biology, immunology, microbiology, and pathology), clinical sciences (surgery, cardiology, dermatology, obstetrics and gynaecology, family medicine, internal medicine, paediatrics, ENT, neurology, orthopaedics, ophthalmology, pulmonology, radiology, rehabilitation medicine, and urology) and behavioural/social sciences (health care economics, epidemiology, health care law, ethics and philosophy, medical psychology, medical sociology, and psychiatry). In Pretoria the September PT was administered 6 months after the start of the academic year (July, table 2), and therefore a correction was applied to the scores of the Maastricht students who wrote the test within one month of studying. The average difference between the total PT score of March and September (equal to the time difference of sixth months with the Pretoria administration) was calculated per year. This constant was derived from the average growth curves of 20 years of progress testing.[19] This procedure was repeated for the three cluster scores. Table 3 presents the percentages added to the individual scores of the Maastricht students to correct for the moment of administration of the September PT.

Subsequently for all three PTs separately four standard scores (z-score) per student were calculated; three cluster z-scores and one z-score for the test in total.[27] Next, per cluster and for the test in total, the mean of these z-scores was calculated for all three tests together per year of training as well as the corresponding 95% confidence interval around the mean. These were graphically depicted. Non-overlapping intervals indicate statistically significant differences.

Table 3.
The percentages added to the individual scores of the Maastricht students to correct for the moment of administration of September PT in Pretoria. Corrections are depicted for the test in total, the basic sciences (BS), the clinical sciences (CS), and the behavioural/social sciences (SC). Data is based upon the historical growth curves and represent the growth of 6 months.[19]

Year	Percentages added for PT scores			
	Total	BS	CS	SC
1	4,03	4,37	3,64	4,49
2	3,69	3,77	3,69	3,60
3	3,35	3,16	3,73	2,72
4	3,00	2,56	3,77	1,83
5	2,66	1,95	3,82	0,94
6	2,32	1,35	3,86	0,05

Results
Table 4 provides an overview of numbers of test questions before and after the culling process, separate for the tests as a whole and for the areas of basic, clinical and behavioural sciences.

Table 4.
Overview of numbers of original and potentially biased items in the three successive PT's in total, and per cluster. Percentages are depicted in italics.

	Total Test		Basic Sciences		Clinical Sciences		Behavioural Sciences	
	n	*%*	n	*%*	n	*%*	n	*%*
Original minus excluded*	757	*100*	320	*100*	316	*100*	121	*100*
Translation problem	70	*9*	28	*9*	35	*11*	7	*6*
Cultural problem	52	*7*	7	*2*	13	*4*	32	*26*
Potentially biased	122	*16*	35	*11*	48	*15*	39	*32*
Remaining	635	*84*	285	*89*	268	*85*	82	*68*

* Excluded after routine post-test review procedures [13]

Approximately one out of six questions were overall identified as potentially biased and were removed. Bias was least found in basic sciences (11%) and most in behavioural sciences (32%). Overall, most bias (9%) was due to (minor) translation errors which alter

the meaning of the question. Most translation problems (11%) occurred in the area of clinical sciences and least in the area of social sciences (6%). Figure 1a presents two examples of potentially biased questions due to a translation problem. Cultural problems occurred in 7% of all items. Least problems (2%) were found in the basic sciences area while most problems occurred in the area of social sciences (26%). Figure 1b presents an example of three questions which are biased due to cultural differences.

Figure 1a.
Two examples of potential biased questions due to translation errors.

*In het rotsbeen lopen een aantal kanalen. Sommigen staan met elkaar in verbinding, anderen niet. Tot deze die **NIET** met elkaar in verbinding staan behoren:*
- canalis facialis en het vestibulum labyrinthi
(juist)

There are a number of canals in the petrous bone. Some are connected with each other and others are not. The following are canals that are connected with each other:
- the facial canal and the vestibulum of the labyrinth
(true)

*Daar is 'n aantal kanale in die os petrius. Sommige is met mekaar in verbinding en ander nie. Die volgende is kanale wat **nie** met mekaar in verbinding is **nie**:*
- canalis facialis en vestibulum labyrinthi.
(waar)

*De navelstreng van een pasgeborene kan één of twee a. umbilicales bevatten. Het risico op een aangeboren afwijking in thorax **of** abdomen is in één van beide situaties groter dan in de andere.*
- Dit risico is groter indien de navelstreng één a. umbilicalis bevat.
(juist)

*The umbilical cord of a newborn can have one or two umbilical arteries. The risk of birth defects of the thorax **and** abdomen is greatest in one of the two situations.*
- the risk is greater if the umbilical cord has one umbilical artery.
(true)

The word OR (of) was translated as AND (en).

Figure 1b.
Three examples of potential biased questions due to cultural problems.

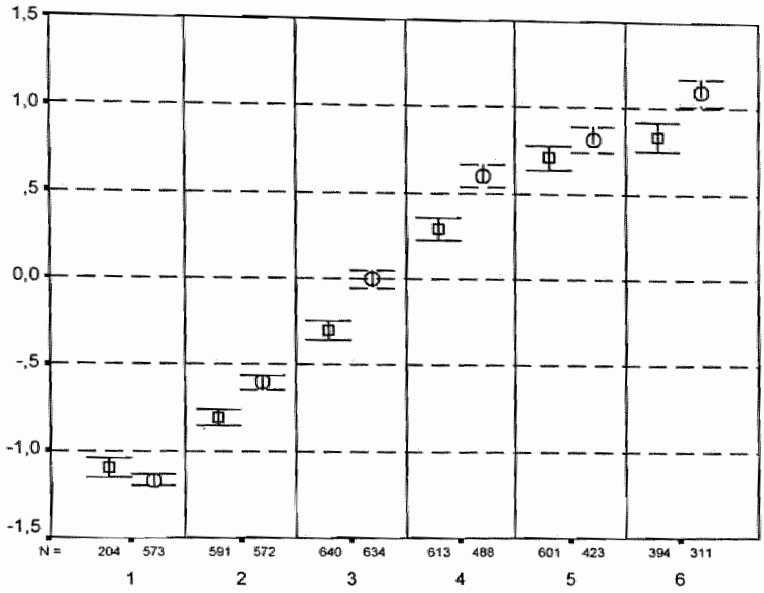
Cystic fibrosis is a hereditary disease. In the Netherlands cystic fibrosis in newborns occurs with a certain frequency.
- The frequency is closer to 1 in 5,000 than 1 in 10,000 newborns.
(true)

The Dutch Act on Working Conditions (ARBO) obligates companies to hire expertise from the ARBO service to carry out certain tasks. These tasks include:
- prescribing medications.
(untrue)

Certain employees may not work underground in a mine.
- This include employees under 18 years of age.
(true)

Figure 2 reports the overall z-score means and confidence intervals of the three PTs graphically. Below the X-axis the number of students that were tested in that class is reported. In general, there is a steady increase of knowledge in both schools across years of training which is astonishingly parallel although the Maastricht students score better in year 2, 3, 4 and 6.

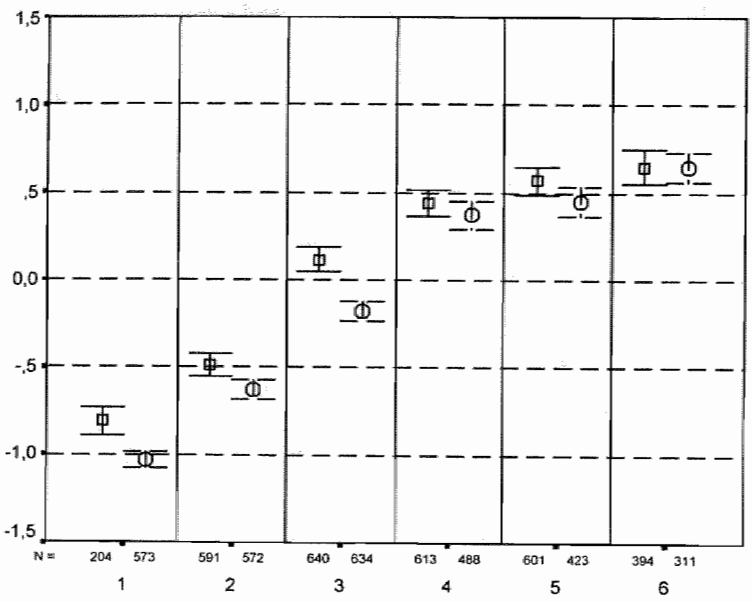
Figure 2.
The students' overall mean percentage correct minus incorrect z-scores (across three tests) and confidence intervals of both Pretoria (□) and Maastricht (○). Below the X-axis the number of students that were tested in that class is reported.



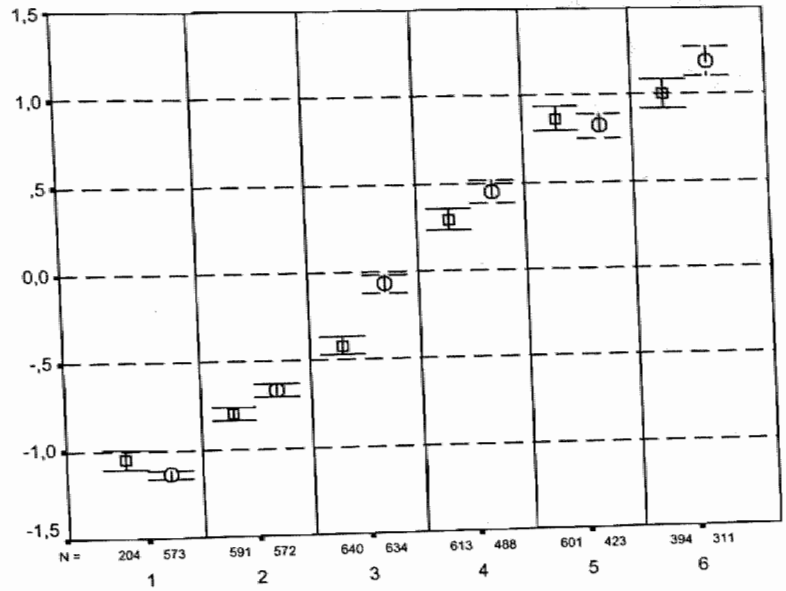
Figures 3a to 3c provide similar information for the basic, clinical and behavioural sciences respectively. The results for the basic sciences seem to indicate that the Pretoria students do better during the first 3 years compared to their colleagues in Maastricht. From Year 4 on forward no difference can be found anymore. The clinical sciences results show slightly higher test scores for the Dutch students in four out of the six years. The tests indicate systematic significant differences for the behavioural sciences between the two schools during all years. The gap seems to diminish in later years, but significant differences remain.

Figure 3.
The students' mean percentage correct minus incorrect z-scores (across three tests) and confidence intervals of both Pretoria (□) and Maastricht (○). Below the X-axis the number of students that were tested in that class is reported. a) Basic Sciences cluster, b) Clinical Sciences cluster, and c) Social Sciences cluster.

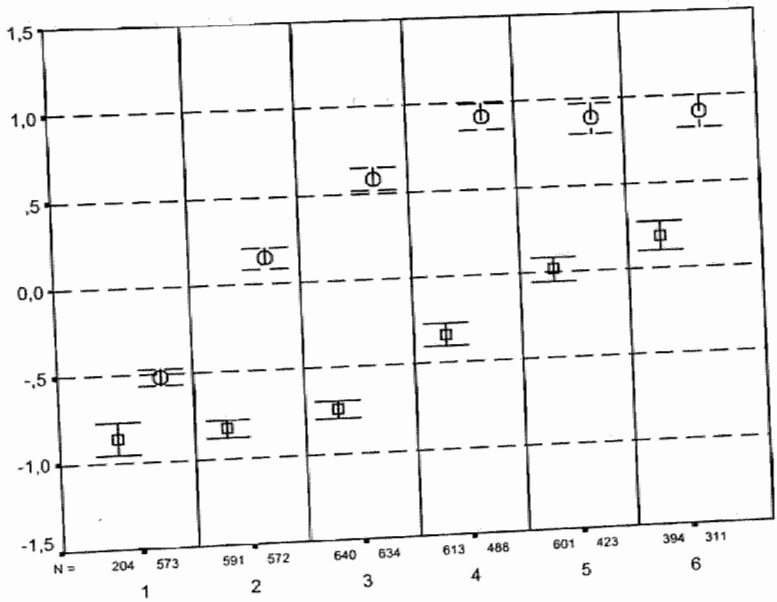
a



b


1,5
1,0
,5
0,0
-,5
-1,0
-1,5
N =
204 573
591 572
640 634
613 488
601 423
394 311
1
2
3
4
5
6

c


1,5
1,0
,5
0,0
-,5
-1,0
-1,5
N =
204 573
591 572
640 634
613 488
601 423
394 311
1
2
3
4
5
6

Discussion

This study has shown that it is viable to translate, administer and interpret progress tests across national boundaries. Approximately one sixth of the questions were potentially biased for cultural differences or due to translation difficulties and errors. Simple translation without further review is thus ill advised and a rigorous review of the translation process is necessary. The variable results found in the Albano et al. study comparing European schools could be the result of the bias in the tests that were used.[23] Nevertheless, sharing item resources across national and cultural boundaries seems worthwhile.

The remainder of test questions that were used for analyses showed interesting outcomes. Overall, the Pretoria medical school seems to perform slightly lower than the Maastricht medical school. Across years of training there is a steady and parallel increase of knowledge. The differences that seem to emerge were threefold. First, the Maastricht students do significantly better on the behavioural sciences. It is noteworthy that significant differences are found in all years. It seems that this finding can be attributed to the accentuation of the Maastricht medical school to the behavioural sciences throughout the entire curriculum. Second, Pretoria students scored better on basic sciences during the first 3 years of medical school. In later years no differences were found anymore. Ultimately, the students seem to have mastered the same basic sciences knowledge. In earlier studies, when PBL and non-PBL programmes were compared, no systematic basic sciences differences were found nationally nor internationally that could be attributed to the integrated PBL approach.[15,28-31] Third, Pretoria students scored slightly poorer on the clinical sciences. Probably this is due to fact that the old curriculum starts predominantly with basic sciences and only later in the curriculum clinical sciences are introduced. Noteworthy is the fact that students, at the moment of this comparison, started their clerkships only in the last year of medical education which could explain the differences found in year six. It seems that the results are a reflection of deliberate accentuations that both curricula seem to pursue: Pretoria has a curriculum with substantial attention to the basic sciences, whereas the Maastricht curriculum has considerable emphasis on public health, ethics, economics, epidemiology, psychology, sociology, et cetera The results on the PT makes these differences well visible.

Methodological considerations

In the Pretoria medical school all students (except first semester first year students and all sixth year students) were expected to participate, but there was no penalty for non-participation and the marks did not count towards anything. The experiment was promoted for its formative, rather than summative, value and the Faculty Curriculum Committee was anxious to be able to compare students outcomes with those of students from a university with an established problem-based learning curriculum. Test results had a purely formative character and no formal assessment consequences. This contrasts with the Maastricht students, for whom the tests were obligatory and part of their certifying exams. Both aspects are essential methodological shortcomings of this study and make a valid interpretation of this comparison difficult.[32] Volunteers in most cases are the better

students.[33] Tests that have no consequences for the participants lead to different answer strategies, particularly more guessing. Indeed, we found that the volunteers (Pretoria) did use significantly fewer questions marks / "I don't know" answers (data not shown). To correct for these different answering strategies, formula scoring was used in this comparison. Another aspect that deserves attention is the difference of familiarity with this kind of test format and formula scoring. The Pretoria students encountered the true-false-do not know format for the first time in July 1998 as the PT is the only test in Pretoria that uses this format and this could influence the results in favour of the Maastricht students. The participation of Pretoria first year students to only one of the three PTs makes the comparison in Year one less reliable. The same, but in a lesser extent is true for Year six. A methodological note should be made concerning the possible influence of the curriculum change that was in progress while the comparison took place. In the light of the results of an earlier comparison of two different curricula in the Netherlands it is not expected to have a significant effect.[15] Supplementary analyses (data not shown) confirm this expectation. However, it can not be ruled out completely. Finally the mathematical correction that had to be made to correct for the time difference that occurred with the administration of one of the three PTs could be a source of error. Although the correction was based on a large data-pool, it remains uncertain whether this average correction was valid for this particular test used in this study. Future comparisons are needed to validate our results.

Conclusion

It is clear that the progress test methodology provides a versatile instrument to assess medical schools across the world. It appears that sharing test material is a viable strategy and that test outcomes are interesting. The sharing of test material is a good strategy for saving resources. Developing good test material, requiring intensive review and effort, is costly, so why not share our resources.[13] A recent international initiative to share item banks across medical schools is quite laudable in this respect (http://www.hkwebmed.org). The cross-institutional comparison of student performance provides good empirical data for debating how we might achieve better standards in medicine at the global level.[34] We could truly achieve a win-win situation by merging economic benefit (the sharing of test material) with quality control (comparing performance across schools).

References

1 Van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher* 1996; 18: 103-9.

2 Van Til CT, Van der Vleuten CPM, Van Berkel HJM. Problem-based learning behavior: The impact of differences in problem-based learning style and activity on students' achievement. *Annual Meeting of the American Educational Research association.* Chicago: 1997. ERIC No. TM026783 (ED409333).

3 Van Berkel HJM, Nuy HJP, Geerlings T. The influence of progress test and block tests on study behaviour. *Instructional Science* 1995; 22: 317-33.

4 Blake JM, Norman GR, Keane DR, Mueller CB, Cunnington J, Didyk N. Introducing progress testing in McMaster university's problem-based medical curriculum: Psychometric properties and effect on learning. *Academic Medicine* 1996; 71: 1002-7.
5 Blake JM, Norman GR, Kinsey E, Smith M. Report card from McMaster: Student evaluation at a problem-based medical school. *Lancet* 1995; 345: 899-902.
6 Verhoeven BH, Van Til CT, Verwijnen GM, Scherpbier AJJA, Van der Vleuten CPM. The consequential validity of the progress test: An investigation into the relationship between test results and problem-based learning behaviour. (under editorial review)
7 Shen L. Progress testing for postgraduate medical education: A four-year experiment of american college of ostepathic surgeons resident examinations. *Advances in Health Sciences Education* 2000; 5: 117-29.
8 Schuwirth LWT, Schrander JPP, Van der Vleuten CPM. Assistententoets kindergeneeskunde: Een beschrijving van de psychometrische eigenschappen. [The national examinaton for residents in pediatrics: A description of its psychometric properties]. *Bulletin Medisch Onderwijs* 1993; 12: 146-51.
9 Van Leeuwen YD. Growth in knowledge of trainees in general practice. Figures on facts [PhD dissertation Maastricht University]. Maastricht: Datawyse, 1995.
10 Ram P. Comprehensive assessment of general practitioners [PhD dissertation Maastricht University]. Maastricht: Unigraphic, 1998.
11 Pollemans M. Kennistoetsing bij huisartsen [Testing knowledge of general practitioners] [PhD dissertation Maastricht University]. Maastricht: Datawyse, 1994.
12 Fourie S, Summers B, Löwes MMJ, Summers RS. Development of a student progress test for the BPharm offered by MEDUNSA, in partnership with TP. *Pharmaciae* 2002; 10: 20-1.
13 Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Schuwirth LWT, Van der Vleuten CPM. Quality assurance in test construction: The approach of a multidisciplinary central test committee. *Education for Health* 1998; 12: 49-60.
14 Verwijnen M, Imbos T, Snellen H, Stalenhoef B, Pollemans M, Van Luyk S, Sprooten M, Van Leeuwen Y, Van der Vleuten C. The evaluation system at the medical school of Maastricht. *Assessment and Evaluation in Higher Education* 1982; 7: 225-44.
15 Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Holdrinet RSG, Oeseburg B, Bulte JA, Van der Vleuten CPM. An analysis of progress test results of PBL and non-PBL students. *Medical Teacher* 1998; 20: 310-6.
16 Bulte JA, Ket P. Forum 1: Invoeren van de maastrichtse voortgangstoets in andere faculteiten: J/O/? [Introducing the Maastricht progress test at other medical schools: True/False/I don't know?]. *Tijdschrift voor Medisch Onderwijs* 2000; 19: 31-7.
17 Verwijnen M, Van der Vleuten C, Imbos T. A comparison of an innovative medical school with traditional schools: An analysis in the cognitive domain. In: Nooman ZM, Schmidt HG, Ezzat ES, editors. *Innovation in medical education: An evaluation of its present status*. Vol. 13. New-York: Springer Publishing Company; 1990: 40-9.
18 Van Hessen PAW, Verwijnen GM. Does problem-based learning provide other knowledge? In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, editors. *Teaching and Assessing Clinical Competence*. Groningen: 1990. 446-51.
19 Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Van der Vleuten CPM. Growth of medical knowledge. *Medical Education* 2002; 36: 711-7.
20 The World Federation for Medical Education. International standards in medical education: Assessment and accreditation of medical schools' - educational programmes. A WFME position paper. *Medical Education* 1998; 32: 549-58.

21 Cohen JJ. Defining international standards in basic medical education: The world federation for medical education has initiated a timely discussion. *Medical Education* 2000; 34: 600-1.
22 Prideaux D, Gordon J. Can global co-operation enhance quality in medical education? Some lessons from an international assessment consortium. *Medical Education* 2002; 36: 404-5.
23 Albano MG, Cavallo F, Hoogenboom R, Magni F, Majoor G, Manenti F, Schuwirth L, Stiegler I, Van der Vleuten C. An international comparison of knowledge levels of medical students: The Maastricht progress test. *Medical Education* 1996; 30: 239-45.
24 Boon JM, Meiring JH, Richards PA. Clinical anatomy as the basis for clinical examination: Development and evaluation of an introduction to clinical examination in a problem-oriented medical curriculum. *Clinical Anatomy* 2002; 15: 45-50.
25 Van der Vleuten CPM, Scherpbier AJJA, Wijnen WHFW, Snellen HAM. Flexibility in learning: A case report on problem-based learning. *International Higher Education* 1996; 2: 17-24.
26 Van der Vleuten CPM. Beyond intuition [Inaugural lecture Maastricht University]. Maastricht: Datawyse, 1996.
27 Norman GR, Streiner DL. *Biostatistics: The bare essentials*. 2nd ed. Hamilton: B.C. Decker Inc., 2000.
28 Vernon DTA, Blake RL. Does problem-based learning work? A meta-analysis of evaluative research. *Academic Medicine* 1993; 68: 550-63.
29 Albanese MA, Mitchell S. Problem-based learning: A review of literature on its outcomes and implementation issues. *Academic Medicine* 1993; 68: 52-81.
30 Berkson L. Problem-based learning: Have the expectations been met? *Academic Medicine* 1993; 68: S79-88.
31 Prince CJAH, Van Mameren H, Hylkema N, Drukker J, Scherpbier AJJA, Van der Vleuten CPM. Does problem-based learning lead to deficiencies in basic science knowledge: An emperical case on anatomy. *Medical Education* 2003; 37: to be published.
32 Schmidt HG. Innovative and conventional curricula compared: What can be said about their effects? In: Nooman ZM, Schmidt HG, Ezzat ES, editors. *Innovation in medical education. An evaluation of its present status*. New York: Springer Publishing Company; 1990: 1-7.
33 Ten Cate TJ. De invloed van anonimiteit op de resultaten van tijdschrijfonderzoek. [The influence of anonymity on the results of time registration research]. *Tijdschrift voor Onderwijsresearch* 1985; 10: 263-73.
34 Mislevy RJ. What can we learn from international assessments? *Educational Evaluation and Policy Analysis* 1995; 17: 419-37.

Chapter 12

Discussion & Conclusions

This dissertation focuses on the utility of progress testing in undergraduate medical education. The model of Van der Vleuten[1] distinguishes five aspects of test utility: reliability, validity, educational impact, acceptability and costs. Validity and educational impact have been addressed by the seven research questions of this dissertation. In this chapter the findings regarding the research questions are discussed and combined with the results of other studies to reach conclusions about the utility of the concept of progress testing.

1. What measures are taken to assure the content validity of the progress test?

The approach of the progress test review committee is described to illustrate the procedure that is used to assure the content validity of the progress test. Four aspects are of key importance: a multidimensional test blueprint, peer review, item analysis and student comments. The blueprint ensures that each test contains a comparable sample of overall medical knowledge. In general, item review by peers leads to adaptation of over 80% of test items (new items and items from previous tests). 97% of the adaptations involve changes in wording and 56% are content-related. The post-test review procedure takes account of item analyses and student comments and results in withdrawal of an average of 5.5% of items and a change of the key of an average 0.7% of items. Pre-test and post-test review identify many serious challenges to content validity, thus improving the quality of the progress test *(chapter three)*.

The study demonstrated that the review procedure led to the detection of flawed items that might compromise the content validity of the progress test. Further research should attempt to corroborate the effectiveness of the different aspects of the review procedure.

2. Does the progress test reflect the final (cognitive) objectives of undergraduate medical education?

In order to answer this question, the content of the progress test was compared to the objectives of undergraduate medical education set out in "Blueprint 1994, training of doctors in the Netherlands".[2] 78% of the subjects of the items of one progress test could be traced in the Blueprint. Considerable differences were found between the three clusters of disciplines included in the progress test, i.e. of the subjects concerning clinical sciences, basic sciences and social/behavioural sciences, 94%, 73% and 47%, respectively were found in the Blueprint. These results should be interpreted with care.caution. Only one progress test was analysed and there are some doubts as to the completeness of the Blueprint, especially in the area of basic and behavioural sciences.[3-5] Comparison between students' test scores on the 179 items represented in the Blueprint and the 51 items not related to the Blueprint revealed striking differences. Sixth year students scored 43% on the related items and 20% on the unrelated items. Scores on the unrelated items remained level after year 2, whereas a normal pattern of improving scores emerged for the related items. In

1988 and 1989, i.e. before the publication of "Blueprint 1994", several studies examined the relevance of progress test questions.[6-9] 554 medical undergraduates (Maastricht medical school 67, other medical schools 487) classified progress test items as relevant, irrelevant or indifferent. The authors found an overall relevance judgement of about 75%, which is comparable to 78% of items that were related to the Blueprint. It is noteworthy that there was considerable variation in students' relevance judgements. Of course, we do not know if the items judged relevant concerned subjects included in "Blueprint 1994".

In all, nearly 80% of the questions in the progress test were found to reflect the final (cognitive) objectives of undergraduate medical education set out in "Blueprint 1994". The scores on these items increased with years of training, unlike the scores on the unrelated items (20%), which were poor. These findings give rise to concern about the content validity of the progress test. It seems that the content validity of the progress test could be improved by using a system of relevance judgement.

3. Is the progress test capable of measuring growth of medical knowledge over successive years of training?

The average overall score on the progress test increases monotonously as a function of training time. Between year 1 and graduation the mean correct minus incorrect score for overall medical knowledge increases from 5% to 41% demonstrating a steady, nearly linear, growth curve. The growth patterns are different for the three science clusters. The curves of the basic and behavioural/social sciences start to level off after 4 years and the curve of clinical sciences is nearly a straight line or even shows a slightly upward curve. The mean scores on basic sciences, behavioural/social sciences and clinical sciences rise from 6% to 38%, from 12% to 37% and from 1% to 44%, respectively. This indicates that parts of the progress test do not optimally use the full range of the scoring scale. There appears to be a ceiling effect for the basic sciences and both a ceiling and a floor effect for the behavioural/social sciences.

Just before graduation, on average more than half of the sixth year students leave 15% of the items in the last progress test unanswered. This may mean that the undergraduate curriculum fails to cover the content of some 15% of test items and that the test does not assess the core content of the curriculum. Another explanation could be that these test items do assess core content but are too difficult or concern subjects not relevant for medical undergraduate education. To differentiate between relevance and difficulty as causes of the unanswered items, for each subject the knowledge level required by "Blueprint 1994" was established by a panel *(chapter four)*. Next, students' test scores and increase in test scores over the years were compared for the different levels of difficulty (according to the "Blueprint"). The absence of correlation suggests that test results and growth in knowledge are more affected by the subject than by the inherent difficulty of an item. This makes sense in view of the self-directed learning environment in which this study was performed. Students determine to what extent a certain topic is studied while the curriculum suggests

which topics are important. This raises another important issue. The "ceiling" effect found for two of the clusters may not be only due to the progress test. Content, structure and format of the curriculum may be equally responsible. The subsiding increase of knowledge at the end of undergraduate medical education for basic and behavioural/social sciences may be due to deficiencies in the planned educational programme, although a shift in students' attention in favour of clinical sciences could also be responsible. The real (hidden) curriculum in year 5 and 6 may be out of step with the official curricular objectives.

From 1974 to 1984 a reference group (50 to 100 persons) of recent graduates from all Dutch medical schools were paid to complete the progress test. The mean scores of sixth year students were in the vicinity of the reference group scores. [10,11] Again in 1988 and 1989 progress test results of a reference group were compared with those of medical students. Some increase in test scores was found after graduation. These results are comparable to earlier findings.[6-9] Bulte et al. used the progress test for evaluating the increase in knowledge during the one-year training programme for general practice[12] in 1988-1989. A significant increase was found on the total test score.[13] These studies suggest that the progress test is able to measure a continued increase in knowledge after graduation. Several other studies also showed the ability of progress tests to measure changes in participants' knowledge level thereby providing evidence for the construct validity of the concept.[12,14-22]

In conclusion, scores on the progress test increase with increasing expertise. The pattern of the increase depends on the content matter tested and the moment in time when students acquire the knowledge. In summary, there appears to be proof of the construct validity of the progress test.

4. Can progress testing be applied in assessing knowledge of skills?

The study that examined this research question was a first step towards longitudinal testing of knowledge of skills in the progress test. It investigated the impact on reliability when a separate written component was added to an OSCE. The 86 items of this knowledge test of skills (KTS) were newly written and the item writers were instructed to write questions that a third year student should be able to answer correctly and at the same time to only address knowledge considered functional for the student upon graduation. The true correlation between KTS and OSCE was 1.00 (observed 0.65). Reliability improved substantially when KTS and OSCE were combined. When the results on a separate knowledge of skills test are added to an OSCE as a separate component the loss in reliability due to the use of fewer OSCE stations can be compensated.

Earlier studies showed that a KTS can be used to measure and compare the increase in knowledge of skills over several undergraduate medical curricula.[23-27] Although these studies were cross-sectional, they showed the ability of a KTS to measure differences in knowledge of skills between students in successive years of training. If knowledge of skills

is tested in a longitudinal manner (i.e. as part of a progress test) it could serve as a screening tool to identify students' weak points in skills performance thus providing useful information for remedial teaching. At the same time a KTS as a component of a progress test can stimulate desired study behaviour in that it encourages students to keep up with knowledge of skills during the year. With only one OSCE at the end of each year students tend to practice skills and study knowledge about skills closely spaced in time. When the KTS is not incorporated in the OSCE, the cognitive aspects of clinical skills will not be overemphasized in preparation for the OSCE and students will use their preparation time to practice skills instead. The summed items on knowledge of skills from the four annual progress tests can be regarded as a KTS and the results can be added as a component to the OSCE.

The concept of progress testing can be and is used for the assessment of knowledge of skills. Since 1999 knowledge of skills items have been included in the progress test. The number of questions about skills per test is still limited and it will probably take another three years before the first students will be longitudinally tested on knowledge of skills during their entire stay in medical school.

5. How reliable and credible is a content-based standard for the progress test?

Two different panels of judges were used to establish a content-based standard for the same progress test through an Angoff procedure. They were asked to set a standard for near graduates. Generalizability theory was used to investigate the reliability as a function of the number of items and judges. Using recently graduated students as judges, the procedure resulted in a reliable, albeit very lenient, standard. Ten graduates rating 200 items would yield a precision (RMSE of 0.51) of the established standard of ± 1% on the scoring scale. Only 7% of the sixth year students would fail when the established cut-off score was applied to the performance data. Using the same procedure, only this time with item writers instead of students, resulted in a rather unreliable and very stringent standard that would have meant that over 55% of students would fail the test. To achieve the same reliability as the graduates yielded with ten judges, 39 item writers would have to judge 200 items. A procedure involving 39 judges is not feasible. Both panels failed to estimate a feasible, reliable and credible standard.

Although the procedures tested in the studies were not suitable for setting a feasible and defensible standard, the results of the studies show that it is essential to take account of test difficulty in setting a standard for progress tests. In both Angoff procedures, most of the variance could be attributed to variation between items. An absolute standard that does not take test difficulty into account will result in major fluctuations in failure rates which are not related to the ability of the candidates. The medical school runs the risk of having to adjust the cut-off scores after test administration when too many students fail the test and revert to using a relative standard.

Clearly, further studies should examine the possibilities of an Angoff standard setting procedure. One should bear in mind that these experiments were only carried out on a part of one progress test and that only 69 sixth year medical students took the exam. Therefore supplementary analyses should be carried out to extrapolate the results to a complete progress test and to calculate cut-off scores for other years using the average growth curves derived from historical data. New procedures with mixed panels and modifications can be used such as not giving the correct answer to the panel members and/or provide p-values (if available). However, the costs involved in item judgement by experts are high. It is advisable to also explore the possibilities of calibrating and/or re-using items to find alternatives for setting a content-based standard by experts.[28-30]

6. Does the progress test have a positive effect on learning behaviour?

Of the three tests used in the undergraduate medical curriculum in Maastricht (unit test, OSCE, and progress test), the progress test score is least correlated with specific test preparation behaviour and most positively correlated with desired learning behaviour. Fewer students engage in test directed study for the progress test (35%) than for the unit test (53%) and the number of hours students claim to spend preparingthemselves for the progress test is insignificant (3) compared to that for the unit test (20) and the OSCE (55). Preparation of the PT is not associated with higher progress test scores. Furthermore, students who reported more deep and active learning behaviour scored higher on the PT whereas low scoring students reported more passive and surface learning behaviour. These results are in line with earlier studies.[17,18,31-34]

There is now strong evidence to support the claim that desired learning behaviour (active self-directed and open-discovery learning) of students during day-to-day learning is not hampered by the progress test.

This conclusion is of major importance since the development and introduction of the concept of progress testing in 1976 was initially motivated by the observation that the tests used at that time had a major negative steering effect on student learning. Students just "studied for the test" in order to increase their chance of passing.[31] Unfortunately, knowledge is subject to rapid decay when it is not used and this particularly happens when knowledge has been memorized just for the occasion.[35-37] For the formation of accurate and long-lasting memory, repetition in learning is indispensable. Closely spaced or massed practice is considerably less effective than practice distributed widely over time. The exact molecular mechanisms underlying these empirical finding are still unknown but recently a protein has been identified that promotes forgetting as a suppressor of learning and memory.[38] Progress testing is a fairly old form of continuous assessment that seems to correspond well to the modern theories of learning and memory.

The absence of a negative influence on learning behaviour and the recent findings in memory research justify a wider use and further development of the progress testing concept within education.

7. Can the concept of progress testing be used to evaluate and compare educational programmes?

Comparisons
The possibilities of the progress test for comparing curricula have been confirmed by a national and an international study. Both studies used a cross-sectional design and provided data of students from all years. To minimise the effect of the specific content of one progress test on cross-sectional comparison, the studies used two and three progress tests, respectively. The comparison between two Dutch medical schools (one PBL and one non-PBL) showed astonishingly small differences, with even the discipline-related scores displaying the same profiles. In the course of the six year curriculum the item scores showed an increasing correlation coefficient indicating that, in the end, many items are answered correctly by both the PBL and the non-PBL students in the Netherlands. The students appeared to acquire the same knowledge, only at different moments in the curriculum. The international comparison used the average test results of three progress tests after removal of items that might cause bias due to cultural differences or translation errors. The results support the decision to do so. On average 16% of the items appeared to be potentially biased. The differences that were found between the test results in the two schools appear to reflect the deliberate accentuations that both curricula pursue.

It appears that the progress test concept is a valid and valuable design for national and international comparisons between curricula. For international comparisons a careful review process should be used to filter out questions that might cause bias due to cultural and epidemiological differences between the participating countries. Additionally, a thorough final check should be performed to correct translation errors. The use of several progress tests is advisable if conclusions are to be drawn at the level of individual disciplines or for small clusters of questions. Otherwise, the sample of interest may be too small to draw valid conclusions.

The utility of the progress test concept

This dissertation focuses on the utility of the progress test as an assessment concept in undergraduate medical education using the model proposed by Van der Vleuten.[1]

$$U = R_{w_r} \times V_{w_v} \times E_{w_e} \times A_{w_a} \times C_{w_c}$$

The main focus of this dissertation is on the validity and educational impact of the concept of progress testing. To be able to determine if the progress test concept is useful, all elements of the model will now be discussed briefly including the three aspects of the model that were not explicitly addressed in this dissertation, namely reliability, acceptability and costs.

Reliability

The overall reliability of the progress test is reasonably good with Cronbach's alphas within years between 0.70 and 0.80 and across years usually above 0.95.[11] Some progress tests have lower reliabilities especially in the first years probably as a result of the limited number of items that reflect the content of the first year of the curriculum. On average, first year students answer 13% of the items on the first progress test and 27% of the items in the fourth test at the end of the year. These findings correspond to the content-based judgement of the teachers in 1989, who found that approximately one out of five items in a progress test reflected the content of the first year of the Maastricht curriculum.[11,39] It is not known whether this percentage has improved since the introduction of the new curriculum.[23,40-42] In practice, three or four tests are used to make a final promotion decision, which improves reliability considerably. On the whole, the reliability of the overall progress test scores is reasonably good but could be improved, especially in Year 1. Improvement might result from sequential testing.[39,43] The normal progress test would be given to all students, and only those students whose scores are too close to the pre-defined cut-off score would answer an additional set of questions thus lengthening the test and increasing reliability. For the other students the test would be over and they would receive their results. Immediate availability of test results would require computerised testing, which is not yet feasible. Another suggestion would be the use of computerized adaptive testing. Shen discusses the use of a one-parameter Item Response Theory (IRT) model (Rasch) to calibrate item difficulty and person ability to avoid "ceiling" or "floor" effects.[19] Indeed, only a limited part of the progress test reflects the content of the first year curriculum causing a "floor" effect. In the IRT procedure a student answers a question, which is followed by a more difficult or easier one, depending on whether the answer was right or wrong. Adaptive testing requires a calibrated item bank in which information of the exact difficulty of items is stored.[44] Furthermore, one of the essentials of the concept is the overview, the X-ray, that is made of every student on the same content at the same time. If this is not possible, many of the advantages of the progress test simply vanish. This means that all content areas must be tested. In order to achieve reliable scores on all disciplines long test sessions are needed,

even with computerized adaptive testing methods. As long as the items are not calibrated computerized adaptive testing is not a realistic alternative. Developing a calibrated item bank is a very resource intensive and expensive project that would quickly exceed the resources of a medical school.

Suggestions
In order to improve reliability in the first year, one might consider adding items that require a lower ability.[39] Overall reliability could be enhanced by using multiple-choice questions instead of true-false questions. Within the same amount of time, a more reliable measurement can be obtained with MCQs.[45]

Validity
Validity is addressed by several studies in this dissertation. It has been shown that the construct validity of the concept of progress testing is excellent. The concept is used in undergraduate and postgraduate (medical) education all over the world and is also applicable in other domains, e.g. knowledge of skills. As described, much time and effort has been invested to ensure the content validity of the progress test. Despite the considerable success of efforts to avoid validity problems, some concerns arise when test content is compared to the final objectives of undergraduate medical education in The Netherlands published in 1994.[2] An important disadvantage of the progress test with regard to criterion validity is the large variation in difficulty of successive tests. This constitutes a major barrier to the use of absolute standards and studies should be carried out to investigate ways to calibrate tests. Both content and criterion validity of the progress test can be improved by using a system of relevance and difficulty judgement.

Suggestions
All staff members involved in item writing and reviewing should take the progress test four times in a row to be better able to conceptualise the target candidates with regard to relevance and difficulty of test items. Recently graduated students should be involved in relevance and difficulty judgement. One might consider appointing one or two recent graduates as members of the progress test review committee.

Educational impact
The favourable outcomes of research on the educational impact of the progress test on individual learning have proven the superiority of the concept in this respect.[17,18,31-34] Evidence was provided that in the context of undergraduate medical education progress testing has a favourable impact on individual learning. The progress test provides a rich source of information for feedback. Students take the test and their results as well as the answer key home with them. Every item is literature referenced and students can use the literature to check why answers are (in)correct. After the post-test review the scores are aggregated to many different sub-scores as described in *chapter three*. Individual as well as average profile scores are sent to the students. Every three months, students are able to ascertain what they do and do not know. Such knowledge is very important to students, residents and qualified medical doctors. Current Dutch legislation presupposes that doctors

are aware of their individual capabilities since a medical doctor is only qualified to practice medicine within his or her individual competencies and capabilities. Also in respect of lifelong learning and CME, it is of vital importance that doctors should be aware of their strengths and weaknesses.[46] Unfortunately, progress test results are almost exclusively used as a decision tool. The detailed information that is provided appears to be used only when remediation is needed. A student poll in 1995 revealed that only 38% of the students used the detailed progress test results to adjust their learning activities.[47] Students should be stimulated to use the results in their daily study activities to correct false assumptions or to spend more time on a topic on which they have repeatedly scored below average. With the introduction of portfolios in the new curriculum in 2002, students are now required to do so. A re-evaluation of the feedback potential of progress testing is therefore in order.

The possibilities offered by the progress test for comparing and evaluating curricula have been shown. Teachers and curriculum designers have access to a rich source of feedback. Items are clustered by discipline to enable direct feedback to the departments that constructed the questions and are responsible for education in that particular discipline. Test results indicate to what extent students have achieved the educational goals of a teaching programme. When the majority of students answer an item incorrectly, this may indicate that a particular topic is addressed inadequately during the course or not covered by textbooks or other learning materials.[48-51] Items that are answered by only a few students (a relatively large percentage of "I do not know" answers) may indicate that the item is not part of the educational objectives of the domain tested or that it was erroneously not included in the programme. Both situations (high percentage incorrect or "I do not know") could indicate the item was poorly formulated. Items that score very poorly in the sixth year are reported to the item writer with the suggestion to remove or adapt the item or adapt the educational programme. However, there are doubts whether action is really taken to evaluate or adapt the content of the curriculum for the subjects that continue to yield poor results. No structural decrease in the percentage "I do not know" answers has been found over the years, indicating that teachers and faculty do not make optimal use of all the information at their disposal.

Notwithstanding the abundant information provided by progress test scores, the feedback potential of the test appears to be far from fully realised. The progress test can and should be used to guide evaluation of weaknesses in the knowledge network of students and detect problems in the curriculum.

Suggestions

Invite item writers, teachers and curriculum designers to discuss progress test results once a year and distil concrete problems and possible solutions. Provide each student with adequate individual feedback and make their results visible by figures and plots (i.e. through a website like the QPE-site[52]). Teach students how to interpret test scores and use them in every day learning activities. Comparing results on unit tests and progress tests can provide unique information about the learning behaviour of individual students.[53] On the basis of the analysis of test results and analyses of learning behaviour through

questionnaires, students failing the progress test can be offered appropriate remedial teaching.

Acceptability

Several polls have been conducted among students since the introduction of the progress test in 1976.[54] Overall student opinion is positive. Question format and relevance are two areas that, over the years, remain less appreciated by the students. Opinions fluctuate over the years and seem mainly to be the result of changes in examination rules. This is especially true for the general opinion about the influence of the progress test on overall promotion. Students would prefer more weight to be given to the unit tests, at the expense of the progress test. This seems logical since students like to be able to prepare themselves maximally and see short time results. They need these results to meet the demands from the government to ensure that they will keep their monthly allowance during their study. This is probably one of the reasons why in the latest poll (2000) 28% (data not published) of students indicated that progress testing should be abandoned in contrast with only 18% in 1995.[47]

Suggestions

The question format should be secondary to the content and as many options should be provided as is realistically possible. Strict use of the "true / false" format should be abandoned and MCQs could be developed. Enrich the items with conceptually and clinically relevant context.[55-58] The questions should approximate the real world as closely as possible. When, eventually, computerized testing becomes a viable option, case-based testing could also be used.[44,56] The purpose and underlying ideas of progress testing should be illustrated to first year students and they should be guided in making use of their test results. One might consider inviting two students to attend the post-test review process as representatives of the students. These suggestions could increase validity, reliability and acceptability.

Costs

The costs involved in progress testing are high. Production of high quality test material is generally tedious and time consuming and progress tests are no exception. Test material has to be constructed with care no matter what assessment concept is used. The extra costs of progress testing are associated with the need for a central organization for test development and the actual test administration itself. The nature of the progress test, i.e. comprehensive and reflecting the end objectives of the curriculum, requires a central organization to develop and organize the tests. A test review committee is costly, but, in our view, essential for item validity. The exact costs are difficult to estimate and depend on the model used.[59,60] Comparisons between centralized and decentralized test organizations are difficult to make. However, if each department would put the same care and effort as the central committees do into developing and organizing their own tests, the costs of a decentralized system would certainly be higher. Cooperation with other medical schools reduces the review and construction costs per school considerably. In the case of the progress test 250 items are constructed four times a year and used to assess about 4000 students at once

(Maastricht: 1340, Nijmegen: 1260, Groningen: 1440). The actual test administration is more expensive because of the large numbers of participating students that have to take the test simultaneously. At the moment there is no easy way to reduce the costs of progress test administration.

Suggestions
Calibrate and expand the item bank in cooperation with other medical schools to make computerized adaptive testing possible.

The utility
The progress test is used in the context of problem-based learning. However, the concept is not limited to problem-based learning as has been illustrated for instance by the Quarterly Profile Test[15,16,61] which was developed independently from the progress test in the same decade. After 25 years, longitudinal assessment methods are now being used in various parts of the world to assess the progress of students in different fields of interest both in PBL and in non-PBL curricula. The name, format, frequency and use of the test results differs per situation. The progress test is but one practical realization of the concept of longitudinal assessment.

The reported studies in this dissertation provide evidence for the utility of the progress test. All elements of the utility model appear to be of good quality. Most importantly, the initial motivation for developing and introducing the concept of progress testing, i.e. to maintain the Maastricht educational philosophy, has been vindicated. Active self-directed and open-discovery day-to-day learning is not hampered by the progress test. Unfortunately, perfect assessment is an illusion and several aspects of the progress test can be improved.

Overall the evidence suggests that the concept of progress testing has a high utility and could make a valuable contribution to assessment in all educational programmes.

References

1 Van der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education* 1996; **1**: 41-67.

2 Metz JCM, Stoelinga GBA, Pels Rijcken - van Erp Taalman Kip EH, Van den Brand - Valkenburg BWM. *Blueprint 1994: Training of doctors in The Netherlands. Objectives of undergraduate medical education*. Nijmegen: University Publication Office, 1994.

3 Metz JC, Bouman LN. Artikelenreeks over actuele ontwikkelingen in het medisch onderwijs [Series of articles on current developments in medical education (editorial)]. *Nederlands Tijdschrift voor Geneeskunde* 1994; **138**: 1014-7.

4 Quaak MJ. Afstemming van artsexamen op het Raamplan 1994. In: Ten Cate TJ, Dikkers JH, Houtkoop E, Pollemans MC, Pols J, Smal JA, editors. *Gezond Onderwijs-5*. Houten / Diegem: 1996. 307-10.

5 Schadé E, Sminia TD. Eindtermen voor de universitaire artsopleiding: "Raamplan 1994 artsopleiding". *Nederlands Tijdschrift voor Geneeskunde* 1995; **139**: 30-5.

6 Van Hessen PA, Verwijnen GM, Imbos TJ. De kennis van de nederlandse basisartsen gemeten met de Maastrichtse voortgangstoets [Knowledge of the Dutch basic physician as assessed with the Maastricht progress test]. *Nederlands Tijdschrift voor Geneeskunde* 1991; **135**: 1975-8.

7 Van Hessen PAW, Verwijnen GM. Does problem-based learning provide other knowledge? In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, editors. *Teaching and Assessing Clinical Competence.* Groningen: 1990. 446-51.

8 Van Hessen PAW. *Referenten bij de voortganstoets van mei 1988.* [PES-publ.nr. 89-06] Maastricht: Universiteit Maastricht, Faculteit der Geneeskunde, 1989.

9 Van Hessen PAW, Verwijnen GM, Imbos T. *Resultaten van nederlandse basisartsen op de Maastrichtse voortgangstoets.* [PES-publ.nr. 89-30] Maastricht: Universiteit Maastricht, Faculteit der Geneeskunde, 1989.

10 Verwijnen M, Imbos T, Snellen H, Stalenhoef B, Pollemans M, Van Luyk S, Sprooten M, Van Leeuwen Y, Van der Vleuten C. The evaluation system at the medical school of Maastricht. *Assessment and Evaluation in Higher Education* 1982; **7**: 225-44.

11 Van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher* 1996; **18**: 103-9.

12 Van Leeuwen YD. Growth in knowledge of trainees in general practice. Figures on facts [PhD dissertation Maastricht University]. Maastricht: Datawyse, 1995.

13 Bulte JA, Verwijnen GM, Tielens VCL, Van Leeuwen YD. Kennis bij huisartsen in opleiding. *Medisch Contact* 1988; **46**: 1426-8.

14 Pollemans M. Kennistoetsing bij huisartsen [Testing knowledge of general practitioners] [PhD dissertation Maastricht University]. Maastricht: Datawyse, 1994.

15 Willoughby TL, Dimond EG, Smull NW. Correlation of quarterly profile examination and national board of medical examiner scores. *Educational and Psychological Measurement* 1977; **37**: 445-9.

16 Arnold L, Willoughby TL. The quarterly profile examination. *Academic Medicine* 1990; **65**: 515-6.

17 Blake JM, Norman GR, Keane DR, Mueller CB, Cunnington J, Didyk N. Introducing progress testing in McMaster university's problem-based medical curriculum: Psychometric properties and effect on learning. *Academic Medicine* 1996; **71**: 1002-7.

18 Blake JM, Norman GR, Kinsey E, Smith M. Report card from McMaster: Student evaluation at a problem-based medical school. *Lancet* 1995; **345**: 899-902.

19 Shen L. Progress testing for postgraduate medical education: A four-year experiment of American college of ostepathic surgeons resident examinations. *Advances in Health Sciences Education* 2000; **5**: 117-29.

20 Tan ES, Imbos T, Does RJMM. A distribution-free approach for comparing growth of knowledge. *Journal of Educational Measurement* 1994; **31**: 51-65.

21 Tan ES, Imbos T, Does RJMM, Theunissen M. An optimal, unbiased classification rule for mastery testing based on longitudinal data. *Educational and Psychological Measurement* 1995; **55**: 595-612.

22 Albers W, Does RJMM, Imbos T, Janssen MPE. A stochastic growth model applied to repeated tets of academic knowledge. *Psychometrika* 1989; **54**: 451-66.

23 Scherpbier AJJA, Verwijnen GM, Schaper N, Dunselman GAJ, Van der Vleuten CPM. Vaardigheidsonderwijs nu en in de toekomst [Current and future skills training]. *Tijdschrift voor Medisch Onderwijs* 2000; **19**: 6-15.

24 Remmen R, Scherpbier A, Van der Vleuten C, Denekens J, Derese A, Hermann I, Hoogenboom R, Kramer A, Van Rossum H, Van Royen P, Bossaert L. Effectiveness of basic clinical skills

training programmes: A cross-sectional comparison of four medical schools. *Medical Education* 2001; **35**: 121-8.
25 Remmen R, Scherpbier A, Denekens J, Derese A, Hermann I, Hoogenboom R, van Der Vleuten C, van Royen P, Bossaert L. Correlation of a written test of skills and a performance based test: A study in two traditional medical schools. *Medical Teacher* 2001; **23**: 29-32.
26 Scherpbier AJJA, Verhoeven BH, Bloemen JCM, Cohen-Schotanus J, Pols J, Rossum HJM, Van der Vleuten CPM, Nieuwenhuizen Kruseman AC. De toename van beheersing van vaardigheden gedurende het curriculum. Een vergelijking tussen Groningen en Maastricht. *Bulletin Medisch Onderwijs* 1997; **16**: 101-7.
27 Scherpbier AJJA. Kwaliteit van vaardigheidsonderwijs gemeten [Assessing the quality of skills training] [PhD dissertation Maastricht University]. Maastricht: Datawyse, 1997.
28 Norcini JJ, Shea JA. The credibility and comparability of standards. *Applied Measurement in Education* 1997; **10**: 39-59.
29 Muijtjens AMM, Hoogenboom RJI, Verwijnen GM, Van der Vleuten CPM. Relative or absolute standards in assessing medical knowledge using progress tests. *Advances in Health Sciences Education* 1998; **3**: 81-7.
30 Embretson SE, Hershberger SL. *The new rules of measurement: What every psyhologist and educator should know.* Mahwah: Lawrence Erlbaum Associates, 1999.
31 Van Berkel HJM, Nuy HJP, Geerlings T. The influence of progress test and block tests on study behaviour. *Instructional Science* 1995; **22**: 317-33.
32 Van Luijk SJ, Melief RM. Toetsresultaten en studiestijlen [Test results and learning styles]. In: Van der Vleuten CPM, Scherpbier AJJA, Pollemans MC, editors. *Gezond onderwijs 1*. Houten / Diegem: Bohn Stafleu Van Loghum; 1992: 111-6.
33 Van Til CT. Voortgang in voortgangstoetsing. Studies naar de aansluiting van de voortgangstoets op probleemgestuurd onderwijs [Progress in progress testing. Studies on the suitability of progress testing within a problem-based educational context] [PhD dissertation Maastricht University]. Wageningen: Ponsen & Looijen, 1998.
34 Van Til CT, Van der Vleuten CPM, Van Berkel HJM. Problem-based learning behavior: The impact of differences in problem-based learning style and activity on students' achievement. *Annual Meeting of the American Educational Research association.* Chicago: 1997. ERIC No. TM026783 (ED409333).
35 Sisson JC, Swartz RD, Wolf FM. Learning, retention and recall of clinical information. *Medical Education 1992*; **26**: 454-61.
36 Semb GB, Ellis JA. Knowledge taught in school: What is remembered? *Review of Educational Research* 1994; **64**: 253-86.
37 Sternberg RJ. Memory processes. In: Klein CP, editor. *Cognitive psychology*. Orlando: Harcourt Brace College Publisher; 1996: 251-78.
38 Genoux D, Haditsch U, Knobloch M, Michalon A, Storm D, Mansuy IM. Protein phosphatase 1 is a molecular constraint on learning and memory. *Nature* 2002; **418**: 970-5.
39 Imbos T. Het gebruik van einddoeltoetsen bij aanvang van de studie [Using assessment of final objectives at the start of a study] [PhD dissertation Rijksuniversiteit Limburg]. Maastricht, 1989.
40 Scherpbier AJJA, Crebolder H, Essed GGM, Van Santen M, Schaper N, Schrander J, Van der Vleuten CPM, Wesseling G, Van den Wildenberg F, Wolfhagen HAP. *Van papier naar patiënt. Gedachten over curriculumverbetering.* Maastricht: Universiteit Maastricht, 1998.
41 Scherpbier AJJA. De Maastrichtse onderwijs benadering [Inaugural lecture Maastricht University]. Maastricht: Datawyse, 2000.

42 Scherpbier AJJA, Crebolder HFJM, Daemen MJAP, Damen JL, Dunselman GAJ, Farla PB, Hillen HFP, Kolle LFJTM, Leiner T, Moulaert VRMP, Nijhuis JG, Oosterhof I, Rosing J, Schrander JJP, Snellen HAM, Snoeckx LHEH, Verwijnen GM, Van der Vleuten CPM, Wesseling GJ. *Voorstel voor het nieuwe Maastrichtse curriculum.* Maastricht: Universitaire Pers Maastricht, 2000.

43 Muijtjens AM, Van Vollenhoven FH, Van Luijk SJ, Van der Vleuten CP. Sequential testing in the assessment of clinical skills. *Academic Medicine* 2000; **75**: 369-73.

44 Kreiter CD, Ferguson K, Gruppen LD. Evaluating the usefulness of computerized adaptive testing for medical in-course assessment. *Academic Medicine* 1999; **74**: 1125-8.

45 Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. *Medical Education* 1985; **19**: 238-47.

46 Abrahamson S, Baron J, Elstein AS, Hammond WP, Holzman GB, Marlow B, Taggart MS, Schulkin J. Continuing medical education for life: Eight principles. *Academic Medicine* 1999; **74**: 1288-94.

47 Brümmer I, Linden ML, Hoogenboom RJI, Scherpbier AJJA, Verwijnen GM. Het oordeel over de Maastrichtse voortgangstoets. Het consumentenoordeel in 1995 [Assessment of the Maastricht progress test. Consumer opinion 1995]. *Bulletin Medisch Onderwijs* 1995; **14**: 174-85.

48 Verwijnen GM. De student als kwaliteitsbewaker. De rol van de student bij de kwaliteitsbewaking van de Maastrichtse voortgangstoets [The student as quality controller. The student's role in quality assurance of the Maastricht progress test]. *Bulletin Medisch Onderwijs* 1994; **13**: 87-95.

49 Prince CJAH, Visser K. The student as quality controller. In: Scherpbier AJJA, Van der Vleuten CPM, Rethans JJ, Van der Steeg AFW, editors. *Advances in Medical Education.* Dordrecht / Boston / London: 1997. 15-8.

50 Visser K, Prince CJAH, Scherpbier AJJA, Van der Vleuten CPM, Verwijnen GM. Students can be full partners in designing their education. *Academic Medicine* 1997; **72**: 1034-5.

51 Verhoeven BH, Verwijnen GM, Scherpbier AJJA, Schuwirth LWT, Van der Vleuten CPM. Quality assurance in test construction: The approach of a multidisciplinary central test committee. *Education for Health* 1998; **12**: 49-60.

52 Anonymous. *Quarterly profile examination.*[Internet site: http://research.med.umkc.edu/qpe/] School of Medicine University of Missouri, 2002 (Accessed: 20 november).

53 Verwijnen GM. Betekenis van studieresultaten bij studiebegeleiding. Een handvol ervaringen [The meaning of test results in supervision. A handful of experiences]. In: De Grave WS, Nuy HJP, editors. *Leren studeren in het hoger onderwijs: Perspectieven voor integratie.* Almere: Versluys Uitgeverij BV; 1987: 95-110.

54 Linden ML, Brümmer I, Hoogenboom RJI, Scherpbier AJJA, Verwijnen GM. De Maastrichtse voortgangstoets. Een vergelijking van drie peilingen van het consumentenoordeel [The Maastricht progress test. A comparison of three samplings of consumer judgement]. *Bulletin Medisch Onderwijs* 1995; **14**: 186-91.

55 Regehr G, Norman GR. Issues in cognitive psychology: Implications for professional education. *Academic Medicine* 1996; **71**: 988-1001.

56 Schuwirth LWT. An approach to the assessment of medical problem solving: Computerised case-based testing [PhD dissertation Maastricht University]. Maastricht: Datawyse Universitaire Pers Maastricht, 1998.

57 Schuwirth LWT, Blackmore DE, Mom E, Van den Wildenberg F, Stoffers HEJH, Van der Vleuten CPM. How to write short cases for assessing problem-solving skills. *Medical Teacher* 1999; **21**: 144-50.

58 Schuwirth LW, Verheggen MM, van der Vleuten CP, Boshuizen HP, Dinant GJ. Do short cases elicit different thinking processes than factual knowledge questions do? *Medical Education* 2001; **35**: 348-56.

59 Reznick RK, Smee S, Baumber JS, Cohen R, Rothman A, Blackmore D, Berard M. Guidelines for estimating the real cost of an objective structured clinical examination. *Academic Medicine* 1993; **68**: 513-7.

60 Clack GB, Baty M, Perrin A, Eddleston AL. The development of a more equitable approach to resource allocation and manpower planning for undergraduate teaching in a UK medical school. *Medical Education* 2001; **35**: 102-9.

61 Willoughby TL, Hutcheson SJ. Edumetric validity of the quarterly profile examination. *Educational and Psychological Measurement* 1978; **38**: 1057-61.

Summary

This dissertation explores the utility of the concept of progress testing in undergraduate medical education. In *chapter one* background information about knowledge, learning and testing is presented together with early difficulties concerning assessment in a problem-based learning environment, which eventually led to the concept of progress testing.

The Maastricht medical school started in 1974 and adopted the instructional method of problem-based learning. Maastricht was the second school after McMaster in Canada to introduce a curriculum that was entirely structured along the lines of problem-based learning. Problem-based learning is assumed to stimulate active, self-directed, lifelong learning. Students play an active role in their own learning: they generate learning issues, decide how to study these issues, and evaluate what they have learned. The students are encouraged to take substantial responsibility for their own learning. Independent and active learning is stimulated by small group discussions. Since the introduction of problem-based learning, test developers and curriculum constructors have been struggling to avoid the pitfall of using tests that steer student learning in an educationally undesirable direction. For the students the examination programme is the real curriculum and if the desired educational objectives are not reflected in and reinforced by the assessment programme, a "hidden curriculum" of assessment based objectives will prevail. Therefore, congruence between educational objectives and assessment programme is of the essence. Educational programmes with stimulation of self-directed learning as a major goal, should enable students to decide, within certain boundaries, what, when and how they will study. In summary, in order to remain faithful to the educational philosophy of problem-based learning, the direct link between the educational programme and assessment had to be severed. It was for this purpose that the progress test was developed. Wijnen, one of the founding fathers of Maastricht University, introduced the concept of progress testing in 1976 and since the 1977-1978 academic year progress tests have been a fixture in the Maastricht Faculty of Medicine's assessment programme. The progress test is best characterized as a comprehensive final examination reflecting the cognitive end-objectives of the undergraduate medical curriculum. It samples knowledge across all disciplines and content areas that are relevant for the medical degree. As a result there is no direct link between the test and any individual course. The progress test is intended to monitor the growth of students' medical knowledge (within the framework of the final objectives of undergraduate medical education). Four times per academic year all students take the same progress test simultaneously. Over the course of the six-year curriculum students sit 24 progress tests. The test results reflect how far students have progressed towards the final objectives of undergraduate medical education.

In *chapter two* a model for evaluating the utility of assessment methods is presented. This model is used to investigate the utility of progress testing. The model relies on five aspects: 1) reliability, 2) validity, 3) educational impact, 4) acceptability, and 5) costs. The utility of any assessment method is determined by a trade-off between the five aspects. The utility of a test depends on the purpose of the test and its properties. Inevitably, the choice of an assessment method will always be the outcome of a compromise between what is desirable

and what is achievable. Two aspects of the utility model (validity and educational impact) have resulted in the seven research questions which are addressed in the following chapters:

- What measures are taken to assure the content validity of the progress test? *(chapter three)*
- Does the progress test reflect the final (cognitive) objectives of undergraduate medical education? *(chapter four)*
- Is the progress test capable of measuring growth of medical knowledge over successive years of training? *(chapter five)*
- Can progress testing be applied in assessing knowledge of skills? *(chapter six)*
- How reliable and credible is a content-based standard for the progress test? *(chapter seven & eight)*
- Does the progress test have a positive influence on learning behaviour? *(chapter nine)*
- Can the concept of progress testing be used to evaluate and compare educational programmes? *(chapter ten & eleven)*

Chapter three describes in detail the construction process and quality assurance procedure of the progress test and how feedback about the test results is provided to students and staff. Four aspects are considered to be of key importance: a multidimensional test blueprint, peer review, item analysis and student comments. The large majority (over 80%) of submitted items are not suitable for inclusion in the progress test without adaptation. The thorough review procedure contributes to the content validity of the progress test although it is difficult to quantify this very precisely. Until 1994, there were no clear, generally accepted, criteria against which the progress test review committee could judge the relevance of test items. The progress test is curriculum independent and the item writer of a department (content expert of that particular discipline) decides which topics are relevant and should be included in a test of the knowledge of a recent medical graduate. Concerns about the relevance of a certain item, whether or not supported by student comment or item analysis, are communicated by the item reviewer to the item writer together with suggestions for change. However, the departmental item writers are under no obligation to take the advice to heart.

In 1994 the objectives of undergraduate medical education in the Netherlands were established in a national consensus procedure and published in "Blueprint 1994". The study described in *chapter four* sought evidence to support the content validity of the progress test by exploring the relationship between test content and the established national objectives. One progress test was analysed and 22% of the core subjects of the test items were not found to be present in "Blueprint 1994". The findings raise some concerns that the content validity of the progress test may be affected by relevance problems.

The concept of progress testing entails measurement of the growth of medical knowledge over the course of the medical curriculum by assessment at regular intervals of the knowledge of students of all classes by comprehensive tests that are not related to any

particular course. *Chapter five* focuses on the construct validity of this concept. Test scores of more than 3000 medical undergraduate students on 84 progress tests were used to calculate the average increase in knowledge during the six-year curriculum. Scores were found to increase monotonously as a function of training time, which supports the construct validity of the total test. Examination of the results for the three main clusters of disciplines included in the progress test revealed three different growth patterns. The curves of the basic and behavioural/social sciences level off after 4 years whereas the curve of clinical sciences continues to curve slightly upwards. This suggests that there is a "ceiling effect" for basic sciences and behavioural/social sciences. This may be associated with the relatively large number of items that are not considered relevant according to "Blueprint 1994", although the content, structure and format of the curriculum could also be responsible. The subsiding increase of knowledge at the end of undergraduate medical education in specific content areas like basic and behavioural/social sciences could be attributable to deficiencies in the planned educational programme but also to a shift of students' attention towards the clinical sciences. The real (hidden) curriculum in year 5 and 6 (clerkships) is possibly out of step with the official curricular (and progress test) objectives. Meanwhile, a new curriculum, implementing more vertical integration, started recently and its effects will have to be awaited.

Further evidence of the construct validity of the concept of progress testing can be gathered by showing its applicability in other domains, like the assessment of clinical skills. In *chapter six* the potential advantages of combining assessment of knowledge of skills with assessment of skills performance (OSCE) are discussed. The study described in this chapter provides proof of the feasibility of this combination from the perspective of reliability. This study has been the first step in the direction of longitudinal testing of knowledge of skills in the progress test. Since 1999 items on knowledge of skills have been included in the progress tests.

Chapter seven and eight focus on the reliability and credibility of content-based standards applied in the progress test concept. The progress test is used in a formative way to inform students of their progress, possible gaps in their knowledge and their relative position vis-à-vis the end-objectives. The test is also used in a summative way to make pass/fail decisions for individual students. The validity of these decisions is dependent on standards. Although the quality control process is very thorough and test construction is highly standardized, consecutive progress tests do not have the same degree of difficulty. This fact hampers the use of absolute standards. For 20 years relative cut-off scores, determined by the overall performance of each class have been used to minimize the number of false positive and false negative decisions. In September 1996 this normative approach was abandoned and a fixed standard based on past test performance across eight years was introduced. This resulted in major fluctuations in failure rates. A possible alternative might be a content-based standard setting procedure, with a panel of experts judging the items. Two Angoff procedures involving different panels that judged the same progress test items have been performed. Both panels were asked to set a standard for near graduates. Generalizability theory was used to investigate the reliability as a function of the number of items and

judges. Using recently graduated students as judges, the procedure resulted in a reliable though very lenient standard. Ten graduates rating 200 items would yield a precision (RMSE of 0.51) of the established standard of ± 1% on the scoring scale. Application of the established cut-off score to performance data of sixth year students resulted in a failure rate of 7%. An identical procedure only this time with item writers as judges resulted in a rather unreliable and very stringent standard which would have caused more than 55% of students to fail this exam. To achieve the same reliability as that obtained with ten recent graduates as judges, 39 item writers would have to judge 200 items. A procedure involving 39 judges is not feasible. Neither the panel of graduates nor the panel of item writers succeeded in estimating a feasible, reliable and credible standard. It is clear that test difficulty should be taken into account, whichever standard setting procedure is used.

The study presented in *chapter nine* sought an answer to the most pressing question with regard to the concept of progress testing. The progress test was developed in order to maintain the educational philosophy of problem-based learning by cutting the direct relationship between specific components of the educational programme and assessment. It was designed to preclude test-directed studying and reinforce self-directed learning. To find out if progress testing has lived up to these expectations, the specific style of problem-based learning behaviour was quantified through a questionnaire and correlated with results on progress tests and two course related tests (unit test and OSCE). Unlike the progress test, the latter two are not designed to preclude test-directed studying nor to prevent students from leaving their individual learning paths. The findings showed that students who report deeper and more active learning behaviour score higher on the progress test and low scoring students claim to have a more passive and superficial learning behaviour. Learning behaviour explains 10% of the variance in progress test scores and only 2% of the variance in scores on the OSCE and the unit test. Test directed preparation does not yield higher progress test scores. On the whole, the results of this study provide evidence that active self-directed and open-discovery learning is not hampered by the progress test. This is in line with the intentions of the progress test.

In *chapter ten and eleven* the progress test itself is used as a source of information for curriculum evaluation. The combination of the validity and the cross-sectional and longitudinal design of the progress test provides a strong evaluation design. The progress test is not tailored to a specific curriculum and therefore can be used to compare different curricula. Several comparisons at macro level were conducted in the past 25 years.

In *chapter ten* the progress tests results of two Dutch medical schools were compared at a detailed level to probe the possibilities and impossibilities of the progress test as an instrument for comparing different curricula. The Maastricht (PBL) and Nijmegen (non-PBL) medical schools were compared. The progress test results showed astonishingly small differences. Even the discipline-related scores demonstrated the same profiles. In the course of the two six year curricula there was an increase in the correlation coefficients of the item-scores, indicating that, in the end, many of the same items are answered correctly by

all students irrespective of the curriculum they attend. The students appear to master the same knowledge only at different moments in the curriculum.

Chapter eleven presents an international comparison of three progress tests. The tests were administered in Pretoria (South-Africa) and Maastricht (The Netherlands). The comparison was conducted after correction for cultural bias and translation problems. On average 16% of the questions appeared to be potentially biased due to translation errors or cultural or societal differences. The differences that were found between the two schools seem a reflection of the deliberate accentuations that the two curricula pursue. Both studies show that the progress test concept is a valid and valuable design to compare curricula nationally as well as internationally.

In *chapter twelve* the results are summarized and the research questions of this dissertation answered. The elements of the utility model not studied in this dissertation (reliability, acceptability and costs) are discussed briefly because of their relevance to overall conclusions regarding the utility of the concept of progress testing. The overall reliability of the progress test is in the range of 0.70 to 0.80 within years and across years it usually is above 0.95 (Cronbach's alpha). This is reasonably good but can be improved, especially in Year 1. Overall student opinion is positive but opinions fluctuate over the years. Question format and relevance are two areas that, over the years, remain less appreciated by the students. The costs involved in progress testing are high but comparisons with the costs of other tests are difficult to make and not available. Collaboration with other medical schools reduces the costs of progress testing considerably. Further research in these three areas is proposed.

Chapter twelve continues with suggestions for possible adaptations of the progress test. Multiple-choice questions could be used instead of the true/false/? format and easier items should be included. The items should be further enriched with conceptually and clinically relevant context and recently graduated students should be involved in test review. All staff members involved in item writing and reviewing should take the progress test to enable conceptualisation of the test difficulty and target candidates. Item writers, teachers and curriculum designers should be invited to discuss the progress test results once a year and distil concrete problems and possible solutions. Each student should not only be provided with adequate individual feedback (as is already being done) but should also be taught how to interpret test scores and make use of that information in every day learning activities. The learning behaviour of students who fail the test should be analysed through questionnaires and remedial teaching should be offered at an earlier stage. Calibration and expansion of the item bank is necessary in order to make computerized adaptive testing possible. This should be done in cooperation with other medical schools to reduce developmental costs. When, eventually, computerized testing becomes a viable option, case-based testing could be implemented to increase authenticity. All these suggestions could potentially increase validity, reliability, acceptability and educational impact but will entail considerable costs, certainly in the short run.

The studies reported in this dissertation provide evidence to support the utility of the progress test concept. Although several suggestions for improvement can be made, all elements of the utility model seem to be of sufficient quality. The reported studies in this dissertation provide evidence for the utility of the progress test. Most importantly, the initial motivation for developing and introducing the concept of progress testing, i.e. to maintain the Maastricht educational philosophy, has been vindicated. Active self-directed and open-discovery day-to-day learning is not hampered by the progress test.
Overall the evidence suggests that the concept of progress testing has a high utility and could make a valuable contribution to assessment in all educational programmes.

Samenvatting

Dit proefschrift onderzoekt de bruikbaarheid van het voortgangstoets concept voor het medisch onderwijs. In *hoofdstuk één* wordt achtergrondinformatie gegeven over kennis, leren en toetsing. Er wordt tevens ingegaan op de eerste moeilijkheden rondom toetsing waar men in de beginperiode van het probleem gestuurd leren tegenaan liep en wat geleid heeft tot het concept van "voortgangstoetsing".

De medische faculteit Maastricht werd in 1974 opgericht en gebruikt sindsdien het Probleem Gestuurd Onderwijs (PGO) als instructiemethode. Maastricht volgde hiermee McMaster in Canada en was de tweede opleiding die geheel volgens de principes van probleem gestuurd leren was opgezet. Probleem gestuurd leren is bedoeld om actief, zelfgestuurd, levenslang leren te stimuleren. Studenten spelen een actieve rol in het eigen leren: zij formuleren zelf leerdoelen, bepalen hoe ze deze onderwerpen bestuderen en evalueren vervolgens wat zij hebben geleerd. De studenten worden aangemoedigd om verantwoordelijkheid te nemen voor het leren. Onafhankelijk en actief leren wordt gestimuleerd door het voeren van groepsdiscussies. Sinds de introductie van PGO worstelen toets- en curriculum ontwikkelaars met het feit dat toetsen studenten vaak in een onderwijskundig verkeerde richting sturen. Het examenprogramma bepaalt het ware curriculum. Als de wenselijke leerdoelen niet worden gereflecteerd en bekrachtigd door het examenprogramma, ontstaat er een verborgen curriculum dat bestaat uit examendoelen aangezien studenten worden "afgerekend" op de behaalde examenresultaten. Het is daarom essentieel dat de leerdoelen van het curriculum en de doelen van het examenprogramma overeenkomen. Onderwijsprogramma's die zelfgestuurd leren hoog in het vaandel hebben staan moeten studenten binnen bepaalde grenzen zelf laten bepalen wat, wanneer en hoe ze studeren. Samenvattend, om trouw te blijven aan de educatieve filosofie van PGO moet de directe koppeling tussen het onderwijs programma en examenprogramma verlaten worden. Dit is de aanleiding geweest voor het ontwikkelen van de voortgangstoets. Wijnen, één van de oprichters van de Universiteit van Maastricht, introduceerde het concept van de voortgangstoets in 1976. Sinds het academische jaar 1977-1978 is de voortgangstoets een vast onderdeel van Maastrichtse toetssysteem. De voortgangstoets kan het best worden omschreven als een alomvattende cognitieve einddoeltoets voor de basisarts. De voortgangtoets test kennis van alle vakgebieden die relevant zijn voor het artsexamen. Er is dus geen directe relatie tussen de toets en een specifiek onderwijsprogramma. De voortgangstoets is bedoeld om de groei van medische kennis van studenten te volgen (binnen het kader van de eindtermen voor het artsexamen). Vier keer per academisch jaar wordt bij alle studenten tegelijkertijd dezelfde voortgangstoets afgenomen. In de loop van de zes jaar durende opleiding tot basisarts maken de studenten dus 24 voortgangstoetsen. De toetsresultaten laten zien in hoeverre er voortgang is richting de einddoelen van het artsexamen.

In *hoofdstuk twee* wordt een model voor de bruikbaarheid van toetsingsmethoden gepresenteerd. Dit model is gebruikt om de bruikbaarheid van de voortgangstoets te testen. Het model is gebaseerd op vijf aspecten: 1) betrouwbaarheid, 2) validiteit, 3) onderwijskundige invloed, 4) aanvaardbaarheid en 5) kosten. De bruikbaarheid van iedere toetsmethode hangt af van het evenwicht tussen deze vijf aspecten. De bruikbaarheid van

een toets hangt zowel af van het doel als de eigenschappen van de toets. Natuurlijk zal de keuze van een toetsmethode altijd een compromis zijn tussen wenselijkheid en haalbaarheid. Twee aspecten van het bruikbaarheidsmodel (validiteit en onderwijskundige invloed) hebben geleid tot zeven onderzoeksvragen die in de volgende hoofdstukken behandeld zullen worden:

Welke maatregelen zijn er genomen om de inhoudsvaliditeit van de voortgangstoets te garanderen? *(hoofdstuk drie)*
Is de voortgangstoets een afspiegeling van de (cognitieve) einddoelen van het medisch onderwijs? *(hoofdstuk vier)*
Is het mogelijk om met de voortgangstoets toename van medische kennis in de loop van de opleiding te meten? *(hoofdstuk vijf)*
Kan de voortgangstoets gebruikt worden bij het toetsen van kennis over vaardigheden? *(hoofdstuk zes)*
Hoe betrouwbaar en geloofwaardig is een inhoudelijke norm voor de voortgangstoets? *(hoofdstuk zeven en acht)*
Heeft de voortgangstoets een positieve invloed op het leergedrag? *(hoofdstuk negen)*
Kan het concept van de voortgangstoetsing gebruikt worden voor het evalueren en vergelijken van opleidingsprogramma's? *(hoofdstuk tien en elf)*

In *hoofdstuk drie* wordt de totstandkoming en de kwaliteitscontrole van de voortgangstoets in detail beschreven. Tevens wordt uitgelegd hoe de toetsresultaten teruggekoppeld worden aan studenten en docenten. Vier aspecten zijn bij de kwaliteitscontrole van eminent belang: multidimensionele blauwdruk, "peer review", item analyse en studentcommentaar. Het overgrote deel (meer dan 80%) van de ingestuurde items zijn zonder aanpassing niet bruikbaar om opgenomen te worden in de voortgangstoets. De grondige review procedure draagt bij tot de verbetering van de inhoudsvaliditeit van de voortgangstoets. Het is echter moeilijk om dit precies te kwantificeren. Tot 1994 bestonden er geen algemeen geaccepteerde criteria waaraan de voortgangstoetsbeoordelingscommissie de relevantie van items kon afmeten. Een inhoudelijke expert binnen een bepaald vakgebied bepaalt zelf welke onderwerpen relevant zijn en opgenomen moeten worden in een toets die de kennis van een pas afgestudeerde basisarts toetst. De expert schrijft een vraag en levert deze met literatuurverwijzing bij de voortgangstoetsbeoordelingscommissie in. De zorg die de beoordelingscommissie heeft ten aanzien van de formulering, inhoud en relevantie van een bepaalde vraag worden met de schrijver besproken. Tevens worden concrete suggesties gedaan voor aanpassingen. De schrijver van de vraag is echter niet verplicht om de adviezen op te volgen. In 1994 werden de einddoelen van de medische opleiding in Nederland vastgelegd middels een nationale consensus procedure. Dit resulteerde in het "Raamplan 1994". De studie die in *hoofdstuk vier* wordt beschreven is uitgevoerd om bewijs te vinden voor de inhoudsvaliditeit van de voortgangstoets. Hiertoe is de relatie tussen de toetsinhoud en de nationaal vastgestelde einddoelen kwantitatief vastgesteld. Na analyse van een voortgangstoets bleek 22% van de onderwerpen van de toetsvragen niet

terug te vinden in het "Raamplan 1994". Dit betekent dat de inhoudsvaliditeit van de voortgangstoets bedreigd wordt door relevantie problemen.

Het concept van voortgangstoetsing behelst de meting van toename van medische kennis gedurende het curriculum door op vaste tijden deze kennis te toetsen bij alle studenten van alle studiejaren middels een alomvattend examen dat los staat van de losse onderdelen van het curriculum. *Hoofdstuk vijf* richt zich op de construct validiteit van dit concept. De toetsscores van meer dan 3000 studenten op 84 voortgangstoetsen werden gebruikt om de gemiddelde toename in kennis gedurende de zesjarige opleiding te berekenen. De voortgangstoetsscore neemt gedurende de opleidingstijd geleidelijk en constant toe. Dit ondersteunt de construct validiteit van de voortgangstoets. De scores op de drie clusters van vakgebieden tonen drie verschillende groeipatronen. De groei in basiswetenschappelijke kennis en gedragswetenschappelijke/sociale kennis vlakt na 4 jaar af terwijl de groei in klinische kennis licht toeneemt tijdens de laatste jaren van de opleiding. Deze bevinding suggereert dat er mogelijk een beperking bestaat in het meetvermogen van de voortgangstoets op het gebied van de basiswetenschappelijke en gedragswetenschappelijke/sociale kennis in de laatste fase van de opleiding. Dit zou te maken kunnen hebben met de inhoud van de voortgangstoetsvragen betreffende deze vakgebieden. Ten opzichte van de klinisch georiënteerde vragen is volgens het "Raamplan 1994" een relatief groter aantal van de vragen betreffende deze clusters irrelevant. Daarnaast zou de inhoud, structuur en vorm van het curriculum voor de gevonden verminderde groei verantwoordelijk kunnen zijn zoals bijvoorbeeld specifieke deficiënties in het geplande opleidingsprogramma. Tevens kan een verschuiving van aandacht en tijd optreden die de studenten hebben en besteden aan bepaalde deelgebieden. Het echte (verborgen) curriculum in het vijfde en zesde jaar (co-assistentschappen) is mogelijk niet in overeenstemming met de officiële (en voortgangstoets-) einddoelen. Ondertussen is een nieuw curriculum ontworpen en opgestart dat meer verticale integratie beoogt. Het effect hiervan op de kennistoename van studenten moet nog worden afgewacht.

Om meer bewijs te verkrijgen voor de constructvaliditeit van het concept van de voortgangstoets is het mogelijk om de toepasbaarheid van het concept binnen een ander domein aan te tonen. In *hoofdstuk zes* worden de potentiële voordelen van de combinatie van een vaardigheidstoets met een schriftelijke kennis over vaardigheden toets besproken. Uit deze studie blijkt dat zo'n combinatie van toetsen met het oog op betrouwbaarheid uitvoerbaar is. Deze studie is de eerste stap in de richting van longitudinaal toetsen van kennis over vaardigheden middels de voortgangstoets geweest. Sinds 1999 worden items betreffende "kennis over vaardigheden" in de voortgangstoets opgenomen.

Hoofdstuk zeven en acht zijn gericht op de betrouwbaarheid en geloofwaardigheid van inhoudelijke vastgestelde zak-slaag grenzen toegepast op het voortgangstoetsconcept. De voortgangstoets wordt op formatieve wijze gebruikt om studenten informatie te geven over hun voortgang, mogelijke lacunes in hun kennis en hun relatieve positie ten opzichte van de einddoelen. De toets wordt echter ook summatief gebruikt om zak/slaag beslissingen te nemen over individuele studenten. De validiteit van zulke beslissingen is afhankelijk van de

gekozen norm en de wijze waarop deze tot stand komt. Ofschoon er een strenge kwaliteitscontrole plaatsvindt en de totstandkoming van de voortgangstoets vergaand gestandaardiseerd is, hebben opeenvolgende toetsen niet dezelfde moeilijkheidsgraad. Dit gegeven bemoeilijkt het gebruik van een absolute norm. Gedurende 20 jaar is daarom gebruik gemaakt van een relatieve norm die afhankelijk was van de gemiddelde prestatie van de jaargroep. Op deze wijze werd getracht het aantal fout positieve en negatieve beslissingen te minimaliseren. In september 1996 werd deze relatieve norm verlaten en een absolute norm geïntroduceerd die berekend was aan de hand van de gemiddelde toetsresultaten van de acht daar aan voorafgaande jaren. Deze wijze van normeren leidde tot grote fluctuaties in het percentage geslaagde kandidaten. Als alternatief zou een norm vastgesteld kunnen worden op basis van de inhoud en dus intrinsieke moeilijkheid van de toets. Hiertoe kan een panel van inhoudsdeskundigen gebruikt worden die de moeilijkheidsgraad van de vragen moet beoordelen. In *hoofdstuk zeven en acht* worden twee Angoff procedures beschreven. Twee verschillende panels hebben dezelfde voortgangstoets beoordeeld en een norm vastgesteld voor bijna afgestudeerde artsen. Het ene panel bestond uit pas afgestudeerde artsen en het nadere uit docenten die vragen aanleverden voor de voortgangstoets. De generaliseerbaarheidstheorie is gebruikt om de betrouwbaarheid te berekenen als functie van het aantal panelleden en beoordeelde toetsvragen. Het panel van de pas afgestudeerde artsen bereikte een betrouwbare maar erg milde norm. Tien panelleden zouden 200 vragen moeten beoordelen om een precisie van ±1% (een RMSE van 0.51) te bereiken. Bij het toepassen van deze zak/slaaggrens zou 7% van het zesde jaar zakken. Een identieke procedure met schrijvers van voortgangstoetsvragen resulteerde in een tamelijk onbetrouwbare en erg strenge norm waarbij meer dan 55% van het zesde jaar zou zakken. Om dezelfde betrouwbaarheid te bereiken als met 10 pas afgestudeerde artsen zouden 39 schrijvers 200 vragen moeten beoordelen. Een Angoff procedure met 39 panelleden is praktisch niet uitvoerbaar. Geen van beide panels kon op een praktisch uitvoerbare wijze tot een betrouwbare en geloofwaardige norm komen. Wel is duidelijk dat, los van de gebruikte methode, de moeilijkheidsgraad van de toets meegenomen moet worden bij het totstandkomen van de zak/slaag grens.

De studie die in *hoofdstuk negen* gepresenteerd wordt, zoekt een antwoord op de meest prangende vraag die met betrekking tot het concept van voortgangstoetsing gesteld was. De voortgangstoets is ontwikkeld om de onderwijskundige filosofie van het Probleem Gestuurd Onderwijs te behouden door de directe relatie tussen specifieke onderdelen van het onderwijskundige programma en toetsing te verbreken. De toets is ontworpen om toetsgericht studeren te ontmoedigen en zelfgestuurd leren te stimuleren. Om na te gaan of de voortgangstoets deze beloften is nagekomen, werd de specifieke leerstijl binnen Probleem Gestuurd Onderwijs gekwantificeerd met behulp van een vragenlijst. De uitkomsten werden vervolgens gecorreleerd met de resultaten op de voortgangstoets en twee programma-afhankelijke toetsen (bloktoets en vaardigheidstoets). In tegenstelling tot de voortgangstoets zijn de bloktoets en vaardigheidstoets niet ontwikkeld om toetsgericht studeren en het verlaten van individuele leerpaden tegen te gaan. De studie toont dat studenten die dieper en actiever leren, hoger scoren op de voortgangstoets. Studenten die

lager scoren op de voortgangstoets beweren passiever en oppervlakkiger te studeren. Het leergedrag verklaart 10% van de variantie in de voortgangstoetsscore en slechts 2% van de variantie in de scores op bloktoetsen en vaardigheidstoetsen. Toetsgericht studeren resulteert niet in hogere voortgangstoetsscores. Over het geheel genomen levert deze studie het bewijs dat actief zelfgestuurd en ontdekkend leren niet gehinderd wordt door de voortgangstoets. Dit is in overeenstemming met de bedoelingen van de voortgangstoets.

In *hoofdstuk tien en elf* wordt de voortgangstoets zelf als bron van informatie gebruikt om het curriculum te evalueren. De voortgangstoets verzamelt per afname informatie over de kennis die studenten gedurende een bepaalde periode hebben opgedaan. Op deze wijze wordt met iedere toets een dwarsdoorsnede gemaakt terwijl er eveneens longitudinaal informatie wordt verzameld. In combinatie met de aangetoonde validiteit is de voortgangstoets bruikbaar als waardevol evaluatie instrument. De toets is niet speciaal toegesneden op een specifiek curriculum en kan daarom gebruikt worden om verschillende curricula met elkaar te vergelijken. In de afgelopen 25 jaar zijn er verschillende vergelijkingen op een macro niveau verricht.

In *hoofdstuk tien* worden de voortgangstoetsresultaten van twee Nederlandse geneeskunde faculteiten op zeer gedetailleerd niveau met elkaar vergeleken om de mogelijkheden en onmogelijkheden van de voortgangstoets als instrument om curricula met elkaar te vergelijken te onderzoeken. De faculteit van Maastricht en Nijmegen werden met elkaar vergeleken. De toetsresultaten toonden verbazingwekkende kleine verschillen. Zelfs de scores per vakgebied toonden overeenkomende profielen. In de loop van de zes jaar nam de correlatie tussen de vraagscores toe. Dit wijst erop dat veel vragen onafhankelijk van het gevolgde curriculum uiteindelijk goed / beter beantwoord worden. De studenten lijken dezelfde kennis tot zich te nemen alleen op een ander moment gedurende de zes jaar.

Hoofdstuk elf presenteert een internationale vergelijking van drie voortgangstoetsen. De toetsen werden afgenomen in Maastricht en Pretoria. De vergelijking werd uitgevoerd na correctie voor culturele bias en taalkundige problemen. Gemiddeld 16% van de vragen bevatte een potentiële bron van bias op basis van vertaalfouten en/of culturele of maatschappelijke verschillen. De verschillen in toetsresultaten van de studenten lijken gebaseerd te zijn op bewuste keuzen die door de curricula ontwerpers zijn gemaakt. Beide studies tonen dat het voortgangstoetsconcept een waardevol en valide ontwerp is om studieprogramma's op nationaal en internationaal niveau te vergelijken.

In *hoofdstuk twaalf* worden de resultaten samengevat en de onderzoeksvragen van dit proefschrift beantwoord. De elementen van het bruikbaarheidsmodel die in dit proefschrift niet worden onderzocht (betrouwbaarheid, aanvaardbaarheid en kosten) worden kort besproken. Binnen jaargroepen ligt de betrouwbaarheid van de voortgangstoets, gemeten met behulp van Cronbach's alpha, tussen de 0.70 en 0.80. Over jaargroepen heen is dit hoger en normaliter boven de 0.95. Dit is een redelijke en acceptabele betrouwbaarheid, maar kan met name in het eerste jaar verbeterd worden. Over het algemeen zijn de studenten positief over de voortgangstoets al fluctueert hun oordeel over de jaren. De

vraagvorm en de relevantie van de vragen zijn twee terugkerende thema's die over de jaren heen door de studenten minder gewaardeerd worden. De kosten van voortgangstoetsing zijn hoog maar een eerlijke vergelijking met de kosten van andere toetsen is moeilijk te maken omdat niet alle gegevens beschikbaar zijn. Samenwerking met andere medische faculteiten kan de kosten per faculteit aanzienlijk reduceren. Voorstellen voor verder onderzoek op bovengenoemde drie gebieden worden in *hoofdstuk twaalf* gedaan. Daarnaast worden suggesties gedaan voor mogelijke aanpassingen van de voortgangstoets. In plaats van juist/onjuist/vraagteken vragen zouden meerkeuzevragen gebruikt kunnen worden om de betrouwbaarheid te verbeteren. Tevens zouden minder moeilijke vragen opgenomen kunnen worden om de betrouwbaarheid in het eerste jaar te laten toenemen. Meer vragen zouden voorzien kunnen worden van klinisch relevante context en de vorm van de vraag zou aangepast kunnen worden aan de inhoud waarbij het aantal antwoordopties kan variëren om zo de aanvaardbaarheid van de toets te verbeteren. Recent afgestudeerde artsen zouden betrokken kunnen worden bij de totstandkoming en beoordeling van de voortgangstoets en alle schrijvers en beoordelaars zouden de voortgangstoets kunnen maken om zich moeilijkheid van de toets beter te kunnen voorstellen. Vragenmakers, beoordelaars, docenten en curriculum ontwikkelaars zouden eenmaal per jaar de voortgangstoetsresultaten kunnen bespreken en hieruit concrete problemen en mogelijke oplossingen destilleren. Studenten moeten niet alleen een gedetailleerde uitslag krijgen, maar ook geleerd worden hoe deze resultaten geïnterpreteerd moeten worden en gebruikt kunnen worden tijdens hun leeractiviteiten. Het leergedrag van studenten die zakken zou middels vragenlijsten geanalyseerd kunnen worden zodat in een vroeg stadium remedial teaching aangeboden kan worden. Uitbreiding en ijking van de vragenbank is noodzakelijk om adaptief toetsen mogelijk te maken. Om de ontwikkelkosten op te kunnen brengen is samenwerking met andere faculteiten noodzakelijk. Als gecomputeriseerde toetsing uiteindelijk mogelijk is, zal het toevoegen van casusgerichte toetsing de authenticiteit van de voortgangstoets verbeteren. Al deze suggesties zouden potentieel kunnen leiden tot een toename van de validiteit, betrouwbaarheid, en aanvaardbaarheid van de voortgangstoets. Tevens zou de onderwijskundige invloed die de toets heeft ten positieve beïnvloed worden. Deze aanpassingen brengen echter, zeker op de korte termijn, aanzienlijke kosten met zich mee.

De resultaten van de in dit proefschrift gepresenteerde studies ondersteunen de bruikbaarheid van het concept "voortgangstoetsing". Ofschoon diverse suggesties voor verbetering gegeven kunnen worden, lijken alle elementen van het bruikbaarheidsmodel van voldoende kwaliteit. Wellicht de belangrijkste bevinding is het feit dat actief zelfgestuurd leren niet gehinderd wordt door de voortgangstoets. Dit was de initiële motivatie om de voortgangstoets te ontwikkelen en te introduceren. Over het geheel genomen lijkt het concept goed bruikbaar in de onderwijspraktijk en kan het een waardevolle bijdrage leveren aan toetsprogramma's binnen onderwijs in het algemeen.

Dankwoord

Het dankwoord en de stellingen worden het meest gelezen. Zoals een recent gepromoveerde collega stelde, doet dit feit het tussenliggende deel groot onrecht aan.[1] Het dankwoord van een proefschrift is echter wel van eminent belang. Ofschoon dit proefschrift het werk is van velen en tot stand is gekomen door de ondersteuning van weer vele anderen blijkt dit niet uit de kaft. Daar prijkt slechts één naam. Het aantal namen dat eigenlijk op de voorpagina zou moeten staan is ontelbaar. Jacques schreef al in zijn dankwoord dat zijn grootste vrees was dat hij iemand vergat.[2] Ik sluit me hierbij aan. In de ruim zes jaren dat ik bezig ben geweest met dit proefschrift hebben talloze mensen op de meest uiteenlopende wijzen een bijdrage geleverd aan dit uiteindelijke product. Graag bedank ik hen allen van harte!

Een dankwoord zou geen fatsoenlijk dankwoord zijn (en ik zou het niet hoeven vrezen) zonder het persoonlijk noemen van een aantal mensen die een bijzondere rol hebben vervuld in de voorbije jaren.

Allereerst mijn ouders. Zij hebben ieder op hun geheel eigen wijze de basis gelegd voor wie ik ben en wat ik doe. Uiteindelijk zal ik een academisch geschoold ambachtsman worden. Ja pap, toch een vak geleerd en ja mam, toch doorgestudeerd. Bedankt voor jullie onvoorwaardelijke steun ook op de (vele?) momenten dat jullie twijfelden aan de koers die ik voer. Pap, dit boek is voor jou omdat ik weet dat je stiekem heel trots zou zijn.

Inge en Rob, wat zou ik allemaal gemist hebben als ik jullie niet als grote broer en zus zou hebben gehad?! Dank jullie voor alle onvergetelijke momenten als kind en puber en het gemak waarmee jullie geaccepteerd hebben dat we vele jaren niet zo vaak bij elkaar geweest zijn door mijn bezigheden met dit proefschrift.

Chief Albertus, jij bent degene die me voor het eerst aan het medisch onderwijs hebt laten ruiken en dit project van de grond hebt getrokken. Tijdens een reis naar Groningen besprak je de mogelijkheden van dit promotie onderzoek en als geen ander wist je me te enthousiasmeren. Ik bewonder je onaflatende energie die je investeert in onderwijs en mensen, de manier waarop je onderzoekers en studenten behandelt als collega's en je uitzonderlijke (politieke) creativiteit. Zonder jou was dit proefschrift niet ontstaan, ons huis in Groningen niet klaar en het boekje nog steeds niet af. Bedankt voor al je bemoeienissen en altijd razendsnelle commentaar.

Makker Maarten, zoals je weet heb ik met regelmaat gedacht dat het moment van het schrijven van een dankwoord niet zou komen. Ofschoon jij toch ook soms je bedenkingen moet hebben gehad heb je me telkens gesteund in mijn keuzes. Als geen ander kun je me een spiegel voorhouden en me laten denken en zeggen wat er zich werkelijk afspeelt. We hebben samen veel gedeeld en ons regelmatig verbaasd over het gevoel van geestesverwantschap. Onze brainstormsessies achter de bureaus tussen de vele boeken terwijl de kinderen rondom ons huiswerk of ruzie maakten waren meer dan inspirerend. Het kostte Cees vervolgens veel overredingskracht om ons op het reeds uitgezette pad terug te leiden. Dank! Dat je mijn paranimf bent is vanzelfsprekend.

Opperhoofd Cees, als een echte schipper was het jouw taak om leiding te geven aan een jonge onderzoeker die alles deed behalve schrijven. Iedere dag nieuwe ideeën, andere analyses, statistiek leren, meer data verzamelen, doceren of met computers in de weer... maar weer geen tekst. Toch kreeg je me aan het schrijven en hoe beroerd het ook was, je wist me positief te bekrachtigen. Als een geboren coach wist je de juiste snaar te raken zonder een onvertogen woord te uitten; motiveren, corrigeren en stimuleren. Deadline na deadline overschreed ik maar jouw geduld bleek eindeloos. De snelheid waarmee je, met je ondertussen beroemde rode pennetje, correcties en suggesties deed is gedurende de zes jaren niet afgenomen. Sinds ik in Groningen woon "doen we het" voornamelijk via e-mail, je mailbox zal voortaan verstoken blijven van gemiddeld 20 onbenullige vragen per dag. Cees bedankt en laten we nu eens eindelijk gaan zeilen.

Wynand, het is een eer om je als mede oprichter van de Universiteit Maastricht en "vader" van de voortgangstoets tot mijn promotoren te mogen rekenen. We hebben enkele boeiende gesprekken gehad over het ontstaan van de Universiteit en de politieke perikelen die daarmee gepaard gingen. Ook herinner ik me een avond bij je thuis dat je Maarten en mij jouw eerste gedachten over einddoeltoetsing bijbracht. Je genereuze steun hebben het mogelijk gemaakt dat ik me als arts onderzoeker kon richten op de voortgangstoets. Nog steeds niet alle vragen rondom voortgangstoetsing zijn met dit proefschrift beantwoord en we moeten eens praten over vervolgonderzoek. Je bijzondere interesse in de Nederlandse taal is me goed bijgebleven, mede omdat ik nog enkele boeken van je heb liggen die ik nodig terug moet geven.

De leden van de leescommissie ben ik zeer erkentelijk voor de tijd en aandacht die zij hebben besteed aan het kritisch lezen en beoordelen van dit proefschrift. Professor von Meyenfeldt wil ik bedanken voor de prettige wijze waarop nog enkele puntjes op de "i" werden gezet. Yvonne, ik denk terug aan de vele uren die we gesproken hebben over het leven en de dood en de teksten die je destijds keer op keer corrigeerde... dank je wel dat je zitting hebt willen nemen in de leescommissie. Arno, statistiek is en blijft een vak. Jij wist sommige geheimen voor me te ontrafelen, maar voor het echte werk moet ik toch echt weer bij je aankloppen. Dank hiervoor. Professor van Schilfgaarde, beste Reinout, het is me een eer en genoegen dat je als mijn opleider zitting hebt willen nemen in de beoordelingscommissie. Niet alleen bedankt voor het beoordelen maar ook voor de tijd en waardering die ik tijdens mijn opleiding Algemene Heelkunde voor dit project kreeg. Professor Zwierstra, beste Rein, onder jouw bezielende begeleiding heb ik destijds, tijdens mijn keuze co-schap Kinderchirurgie, de keuze gemaakt om te solliciteren voor de opleiding Algemene Heelkunde. Je bent terecht beroemd als chirurg en docent en voor mij functioneer je voor beide als een positief voorbeeld. Dank!

De Skillbillies
Ellen, mijn kamergenoot op het Skillslab van het eerste uur. Hard werken en veel overleg... over zaken en privé. Wat hebben wij gelachen, gehuild, gevloekt en gemopperd. Ik denk nog vaak terug aan die tijd! Mereke, jij als noorderling in Maastricht en ik als zuiderling in

Groningen, de oostelijke as is virtueel versterkt door onze onophoudelijk e-mail contacten. Dank voor alle teksten die je secuur en snel corrigeerde of vertaalde. Katinka, behalve dat ik je dankbaar ben voor alle klusjes die je voor me wilde doen als ik weer eens "iets moest" in het zuiden terwijl ik zelf in het noorden zat, blijf ik je het meest dankbaar voor je rol als koppelaar en de ondersteuning die je aan Cis gegeven hebt. Zoen.

Jan, Hieke, Pie, Els, Toon, Marjo B., Marjo F., Bert, Hanneke, Sophie, Juul, Jocelyn, Margje, Gijs, Martin, Jan (Disaster Day) Joost, Jean, Rina, Roger, Emer, Lidewij, Marleen, Esther...ofschoon we elkaar slechts zelden zagen (zeker de laatste jaren), was het goed toeven temidden van jullie allen. Het Skillslab was en is een heerlijke werkplek!

Robert en Ruud, dank voor alle computer en netwerkondersteuning. Ron bedankt voor de vele SPSS lessen en de eindeloze databrij van voortgangstoetsuitslagen die je me keer op keer met een druk op de knop kon bezorgen. Hetty, Ineke, Diana, Arno, Cita, Marianne, Resie... vaak ben ik onaangekondigd binnengelopen voor raad, advies of een praatje. Jullie deuren stonden altijd open. Dank hiervoor. Lilian, dank je voor al het werk achter de schermen... ik wist niet hoe ik een promotie moest regelen... en nu weet ik het nog niet. En dat dankzij jou! Veel dank.

Lambert, mien jong, wat kan ik ervan zeggen. Je hebt me geweldig bijgestaan als collega, super VBC-er, mede onderzoeker en criticus. Ik mis de wetenschappelijke en hersenkrakende discussie maar ook de vele schaterpartijen. Judith, als onvermoeibare rokende secretaris heb je me de VBC tijd doorgeloodst. Wat moesten ik en de VAX zonder jou? (Ex-) leden van de VBC, dank voor alle inspanningen om de voortgangstoets te maken en te verbeteren tijdens de vele memorabele vergaderingen. Mijn tinnen bord koester ik.

Collega's uit het AZG en MCl, staf, chivo's, fellows en collega-assistenten, dank voor de opvang van een onervaren zuiderling en de kansen die jullie me geboden hebben om werk, opleiding, onderwijs en promotie te combineren. Herbert, het geeft me een veilig gevoel een brede collega en ex-commando naast me te hebben staan. Op naar jouw ceremonie!

Marianne, Ernst, Eva, Marije, Laura, Ewout en Jelle het waren bijzondere tijden waarin Cis en ik vaak bij jullie achter de tafel aangeschoven hebben. Heerlijk hoe dat kon en ging. Er valt veel meer te zeggen maar het is fijner om het gewoon nooit te vergeten.
Fred en Jan-Henk... gelukkig ontvoerden jullie me soms en namen me mee naar zee en zoute whisky. De tijden worden beter! Ofschoon veel van mijn vrienden zich ondertussen wel zullen afvragen of ik wel een telefoon kan bedienen en adressen bewaar, zijn ze trouw gebleven. Gelukkig kennen ze me goed genoeg om te weten dat ik ooit weer boven water kom en plotseling voor hun neus sta. Jeroen, Jacques en Esther, dank voor jullie eindeloze geduld met me. Carien & Patrick, jullie hebben de twijfelachtige eer gehad om van zeer dichtbij het wel en wee rondom dit proefschrift te mogen volgen. Dank jullie wel voor jullie onstuitbare steun en vriendschap.

"Ik wil mijn naam niet eens in dat *&#boekje van je". Cis, al 12½ jaar steunen we elkaar op onze bijzondere manier. We schaven elkaar bij, niet zonder horten of stoten, maar wel vol overgave. Je hebt meer dan alles gegeven om mij deze klus te laten klaren. Zonder jou was dit nog niet af geweest! Ik kan met niets beschrijven wat voor een tijd en energie het jou gekost heeft om dit proefschrift tot een einde te zien komen. Simpelweg bedanken is hier niet op z'n plaats en gelukkig zul je me dat wel helpen herinneren. En al ben ik heel erg blij, je hebt vast gelijk en bent nog blijer dan ik. Het is af! "Bijna" bestaat niet meer. Samen met Daan gaan we genieten en jouw proefschrift tot een goed einde brengen. Daan, jouw naam zou hier eigenlijk niet eens mogen staan. Na 16 maanden is er nu tijd voor pappa dingen!

Literatuur

1 Sonneveld DJA. Testicular germ cell tumours. New insights in epidemiology, genetic susceptibility and outcome [PhD dissertation Rijksuniversiteit Groningen]. Groningen: De Regenboog, 2002.
2 Maas JWM. Invasion and angiogenesis in endometriosis. Experimental studies in the chick embryo chorioallantoic membrane [PhD dissertation Universiteit Maastricht]. Maastricht: Datawyse, 2001.

Curriculum Vitae

Bas Henk Verhoeven is geboren op 10 januari 1970 te Oisterwijk. In 1988 voltooide hij zijn VWO opleiding aan het Maurick College te Vught. In datzelfde jaar startte hij met de studie geneeskunde aan de Universiteit Maastricht (destijds: Rijksuniversiteit Limburg). In 1992 liep hij stage in het Princess Alice Hospice in Esher (Groot-Brittannië). Van 1991 tot 1993 was hij als student-assistent werkzaam bij de afdeling pathologie in Maastricht. Van 1993 tot 1995 was hij lid van de Opleidingscommissie van de Faculteit der Geneeskunde. In 1994 raakte hij betrokken bij onderzoek van medisch onderwijs en verrichte hij samen met vele anderen onderzoek naar de inhoud van het co-assistentschap Gynaecologie/Obstetrie en de relevantie van de voortgangstoetsvragen. In 1995 werkte hij als keuze co-assistent bij de afdeling Kinderchirurgie in Groningen. In 1996 is hij afgestudeerd en in eerste instantie als docent/onderzoeker aangesteld bij het Skillslab in Maastricht. In datzelfde jaar werd hij door onvoorziene omstandigheden ad interim voorzitter van de voortgangstoetsbeoordelingscommissie en startte hij met het onderzoek dat leidde tot dit proefschrift. In 1998 verhuisde hij naar Groningen en begon met de opleiding Algemene Heelkunde onder Professor R. van Schilfgaarde en Professor H.J. Ten Duijs. Sinds 1999 is hij betrokken bij het Gestructureerd Cursorisch Onderwijs dat georganiseerd wordt voor de opleidingsassistenten Heelkunde. Sinds december 2002 is hij werkzaam in het Medisch Centrum Leeuwarden waar onder dr. W.J.H.J. Meijerink de opleiding wordt voortgezet. Bas woont sinds 1993 samen met Ciska Buijs. Zij hebben een zoon van 18 maanden: Daan.